EVOLUTION OF BORN-DIGITAL MOVING IMAGE PROCESSING

Moving to scalable and sustainable workflows

Rachel Curtis
Library of Congress
USA
rcur@loc.gov

Laura Drake Davis
Library of Congress
USA
ladavis@loc.gov
ORCID 0000-0001-9892-2932

Abstract – Long-term preservation of born-digital moving image content is similar to that of any other file-based content in many ways. However, large file sizes, specialized equipment and resources, significant processing storage needs, and the movement of large files are challenges to creating sustainable and scalable workflows. The Moving Image Section of the National Audio Visual Conservation Center at the Library of Congress is making great strides in the development of sustainable and scalable workflows through an understanding of the technical infrastructure, moving image file characteristics and requirements, and the adoption of automated workflows using a combination of open source software and hardware resources.

Keywords – moving image, digital workflows, scalability, technical infrastructure, digital preservation

Conference Topics – From Theory to Practice; Immersive Information.

I. INTRODUCTION

The National Audio Visual Conservation Center (NAVCC) at the Library of Congress (the Library) is home to the world's largest collection of moving image and audio materials. NAVCC, as at other institutions, is experiencing a shift from analog to born-digital, and participates in broader efforts at the Library to establish a community of practice. However, the characteristics of born-digital moving image files present challenges in terms of file size, processing resources, storage allocations and network bandwidth. Over the last ten years, NAVCC staff have worked to address these challenges and anticipate future needs for born-digital moving image processing.

This paper discusses the evolution of born-digital moving image processing workflows, the impact of ongoing IT modernization efforts, resulting challenges in adapting to new internal requirements, and the efforts to ensure workflow sustainability when met with increased numbers of born-digital files.

II. INITIAL BORN-DIGITAL MOVING IMAGE PROCESSING EFFORTS

The Moving Image Section's first born-digital moving image collection workflow was developed for The HistoryMakers Collection. This significant collection consists of oral histories of prominent African Americans from a wide range of disciplines. The ingest of born-digital files was a new endeavor for the NAVCC, but a workflow was adapted from the existing digitization workflow, including verifying checksums, generating derivative files, creating ingest documents, and linking files to their corresponding metadata records.

As illustrated in Figure 1, the initial processing workflow for this project consisted of two parallel paths that converged before ingest. The NAVCC Moving Image Processing Unit created local

iPRES 2023: The 19th International Conference on Digital Preservation, Champaign-Urbana, IL, US.

https://hdl.handle.net/2142/121129

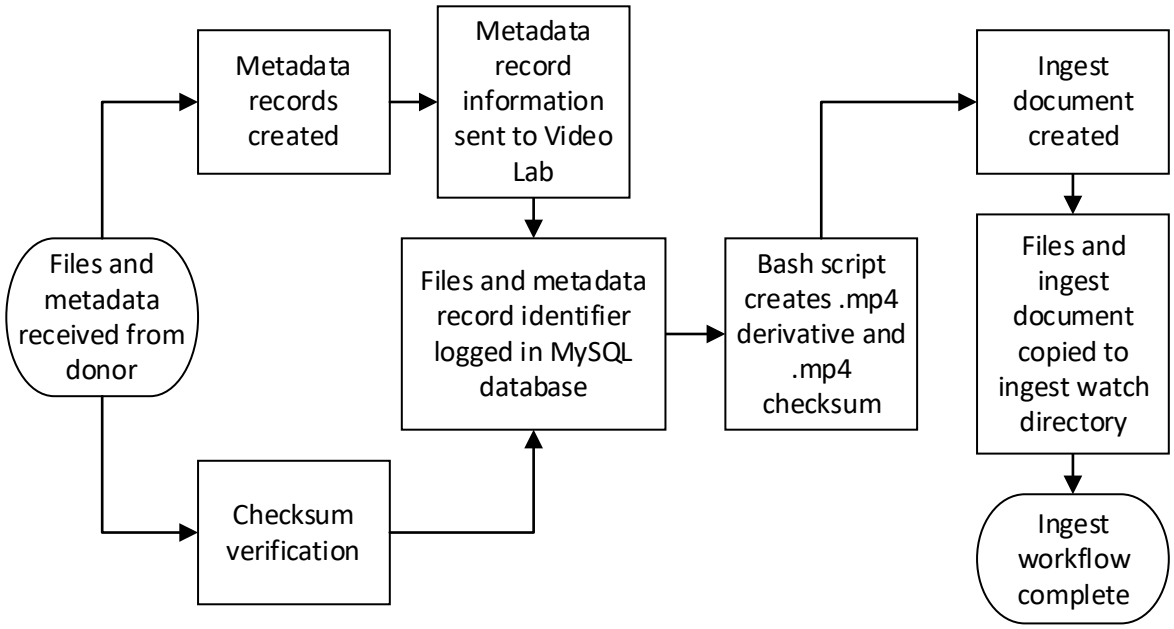

Figure 1: Initial born-digital moving image processing workflow

metadata records and Library of Congress Name Authority records; and the Video Lab created derivative files (.mp4) using OpenCube software [1] and ingested the preservation and derivative files. The metadata creation processes benefited from extensive information from The HistoryMakers organization including interview dates and interviewee biographical information.

This project demonstrated the need for staff dedicated to processing born-digital collections to ensure sustainability and the scalability. The initial workflow relied on the availability of the head of the Video Lab initiate and monitor the modified digitized workflow while balancing day-to-day Video Lab responsibilities. However, it would not be until the establishment of the American Archive of Public Broadcasting (AAPB) that a digital project specialist was hired and dedicated to born-digital collections.

III. A Born-Digital Program Begins

The AAPB began as a project funded by the Corporation for Public Broadcasting (CPB). In 2010, CPB conducted an inventory project and provided funds for 100 public television and radio stations to digitize items in their collection, which resulted in the creation of about 73,000 files. In 2012, CPB selected the Library and the public media station GBH to be the co-stewards of the archive. In this collaborative partnership, the Library is the preservation arm of the archive, ingesting high-resolution preservation files and ensuring their preservation for generations to come, while GBH makes files accessible on the AAPB website. In 2013, the Library hired a limited-term digital project specialist assigned to the AAPB with CPB funding.

The Library received 73,000 files on LTO tape in 2015. The files were delivered according to the BagIt specification [2] along with a master spreadsheet that contained filenames, metadata, and LTO barcodes after the vendor completed digitization. To facilitate the immense job of ingesting these files quickly, NAVCC staff adapted and expanded the scripts developed for HistoryMakers.

This initial AAPB workflow consisted of the following steps:

- Create a SQL database to store all datapoints
- Verify checksums in the bags
- Create a metadata record in the Library's MAVIS system for each file
- Move the media file and any sidecar files (such as .srt files) from the bags to a watch folder
- Create the ingest package for each set of files
- Ingest the files

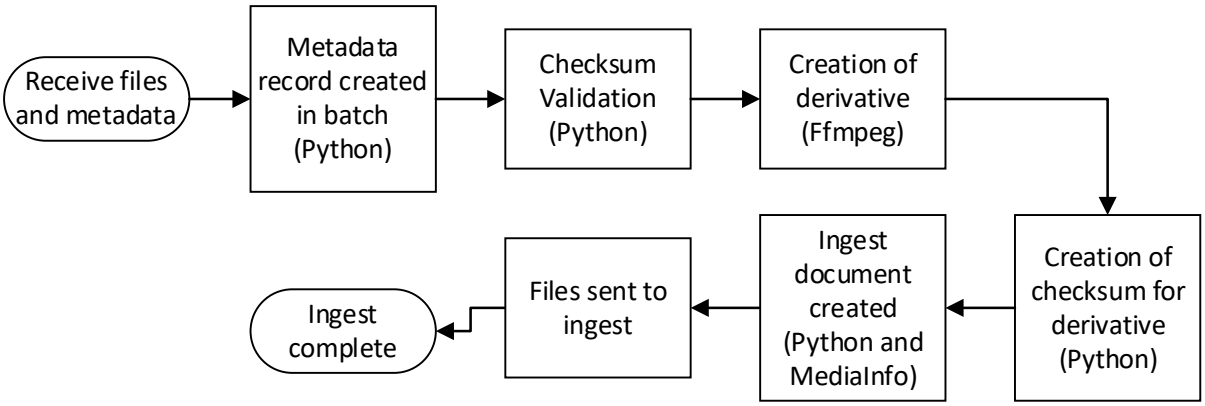

Figure 2: Hybrid manual and automated workflow

Checksum verification occurred at both the ingest package creation stage and ingest stage. The MySQL database [3] then stored the checksums, MAVIS ID, and a timestamp for when each step was completed.

The development and management of this workflow was distributed among staff members from a variety of functional and administrative areas. At the time, there was no dedicated staff member assigned to develop the workflow with integration into the NAVCC systems. The CPB-funded position was focused on overall project management and resolving any issues with the files as they were reported by the video lab supervisor.

In 2015, the Library hired a permanent AAPB digital project specialist responsible for the development and maintenance of the AAPB workflows. The digital project specialist quickly implemented changes to the initial workflow to accommodate new files being received through the AAPB project. This included shifting delivery from LTO-tape to hard drive, requiring pilot batches, adding a quality-control component, and requesting monthly file delivery rather than receiving all files upon a project's conclusion. The shift to hard drive from LTO was deemed necessary due to issues experienced with the LTO drive. Requiring pilot batches and running files through a basic QC profile in the Library's QC software, Baton [4], allowed both the Library and partner institutions to identify issues earlier in the process and relay these issues to the vendor in a timely manner. A cap of 25,000 files accepted per year was also implemented to prevent the accumulation of a backlog.

IV. Scalability, IT Modernization, and Challenges

As born-digital acquisitions increased, the Library hired two additional permanent digital project specialists in 2016 to process and ingest born-digital material outside the AAPB collections - one in the Moving Image Section, and one in the Recorded Sound Section. These two digital project specialists are devoted to born-digital collection work within their respective sections while also sharing and gaining insight from one another, with the goal of creating efficient processing workflows.

To meet this goal, the digital project specialists adapt to an ever-evolving technical infrastructure, modify workflows, and advocate for local technology updates based on observation, experience, and analysis of incoming collections. However, library-wide hardware and software changes are particularly challenging as the NAVCC workflows utilize different processes and systems than the rest of the Library. To tackle such constraints, the digital project specialists must often take the lead in resolving hardware and software changes and providing recommended solutions to improve and enhance current workflows.

Once the basic processing workflows (see Figure 2: *Hybrid manual and automated workflow)* were established for both the AAPB and general Moving Image collections, the staff at NAVCC began to investigate modernizing workflows for specific

https://hdl.handle.net/2142/121129

collections by implementing scalable, automated workflows. For an automated workflow to be successful, it should meet some minimum criteria: 1) consistently formatted machine readable metadata; and 2) consistent file naming and delivery.

Using a series of Python scripts and open source software such as mySQL, FFmpeg [5] and MediaInfo [6], automated workflows were developed based on individual processes found in the early processing workflows. Metadata records were also incorporated into the ingest packages for select collections if sufficient metadata was readily available. The next generation of this workflow evolution will include Baton quality control software [6] and a Dalet AmberFin transcoder [7] to replace FFmpeg during the creation of .mp4 files. The AmberFin is a shared resource at NAVCC with six servers and two transcoding engines per server, providing faster derivative and checksum creation for processing.

Each workflow generation brings unique challenges. Ultimately, a balance must be struck between multiple simultaneous workflows, the amount of available processing resources in the shared environment, and storage, and network bandwidth capabilities to write and move files.

At NAVCC, the evolution of digital moving image formats is ongoing – particularly related to files received from the entertainment industry – and we are prepared for these changes. The shift from SD to HD in television increased the file size by anywhere from 60% to 450% per hour, based on individual file characteristics, and we will see another significant increase in file size with the adoption of 8K resolutions. Increased file sizes create significant challenges in transferring files. For example, in 2017, most born-digital collections were sent via external hard drive to the Library. Currently, some collections are still received on a hard drive, some via SFTP, and still others are received via Amazon Web Services (AWS) or Aspera, a common entertainment-industry file transfer application. However, all these transfer mechanisms come with their own difficulties – hard drives require on-demand virus scanning and lengthy off-load time; SFTP transfers require IT department intervention; and Aspera transfers require navigating a rigorous security process that can take months to complete. Yet, despite these issues, the Moving Image Section is moving towards receiving more collections via SFTP and cloud transfer.

Processing storage and network bandwidth are also factors the Library must consider when attempting to increase our digital file transfer receipts. As noted previously, file sizes are increasing exponentially - a recent acquisition of a 4K motion picture was 781 GB for a 2-hour title - an average of 390 GB per hour. Conversely, processing storage is limited, relying on a constant movement of files during processing and ingest activities to remain viable. Downtime of systems and processing resources quickly result in an accumulation of files, consuming vital storage space. While the storage allocated for processing is generous - 80TB for AAPB and 100TB for other collections - large files require immense storage space. In this respect, we have more control as to what is in the processing space with hard drive transfers than with direct file transfers that are routed to specific directories for processing. If the processing space is full, we can opt not to offload an external hard drive, but direct file transfers will be received if there is available space. While some SFTP transfers occur overnight or in the early morning hours to minimize impact on overall network bandwidth, transfer systems such as Aspera utilize a “pull” wherein files are manually requested from an external source, often occurring during regular business hours, which are typically a peak network period.

Navigating current and changing infrastructure at the Library and NAVCC is a large component of the digital project specialists’ responsibilities. From advocating for infrastructure changes to navigating IT modernization and support issues, the digital project specialists work to maintain current workflows and optimize these workflows for future scalability.

V. The Current Born-Digital Moving Image Workflow Model

Over the past year, several Library-supported born-digital projects have seen an increase in file delivery and are planning to scale up even further. Scalability is critical for moving image processing to ensure maximized processing and leveraging of available resources to meet processing goals. The result is a flexible, scalable workflow model that can be adapted based on the characteristics of each collection (see Figure 3).

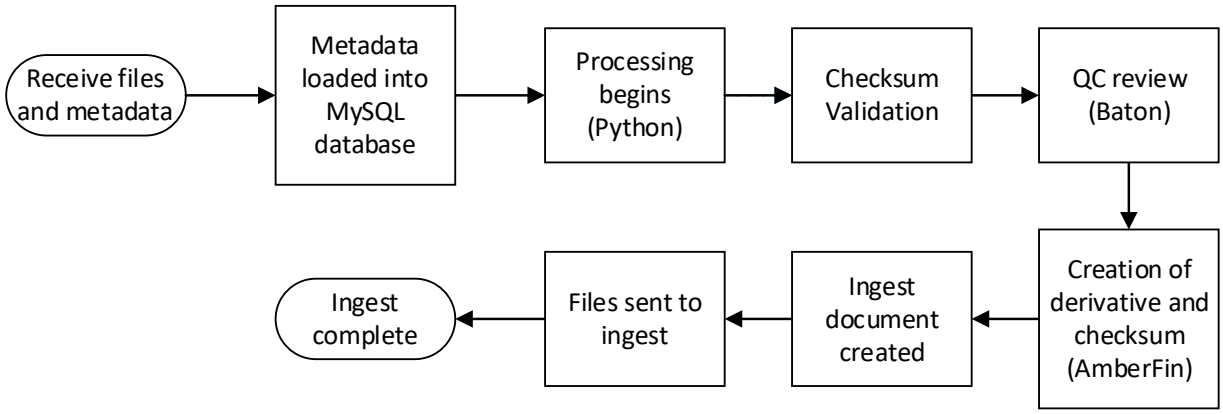

Figure 3: Current flexible and scalable born-digital moving image processing workflow model

Various tools, including Python scripting, MediaInfo, FFmpeg, OpenRefine and some limited direct file deposits, have been incorporated into AAPB workflows to prevent a backlog. This has resulted in two basic workflows for AAPB. The first workflow has manual elements to accommodate file deliveries without checksums and standardized file names. The second workflow is highly automated, and used for files delivered from a vendor, building on the flexible workflow model in Figure 3.

The Congressional Video project is another example of the desire to receive more content through sustainable scaling. The Library receives moving image files from the U.S. House of Representatives and U.S. Senate and is looking to expand content received, ensuring complete overlap with the National Archives and Records Administration. To do so, the Moving Image Section is working with the U.S. House of Representatives Recording Studio and U.S. Senate Recording Studio to standardize file delivery and file formats within the technological abilities and preferences of all project partners.

The U.S. Senate Recording Studio transitioned to solely file-based recordings in 2008. The Library began receiving daily file transfers in 2016, creating the impetus to develop the first automated processing workflow. Using a combination of Python, MediaInfo, FFmpeg, and a mySQL database as well as metadata provided by the Senate, this workflow validates file integrity via checksum verification, creates an .mp4 derivative file, issues the .mp4 checksum, gathers duration information, writes the metadata record, and generates the ingest file. This workflow has been in production since 2018 and is the foundation for automated workflows for other collections. The Moving Image Section is increasing the number of workflows that utilize these elements to enhance performance and allow the digital project specialists time to spend on projects that do not meet the requirements for automated workflows.

The extent of scalability for the non-AAPB collections is undergoing testing with the addition of the Congressional collections. Currently, each collection workflow uses its own virtual machine (VM), mostly due to the local transcoding function. However, with transcoding activities being moved to the Dalet AmberFin transcoder so processor-intensive work can be completed outside of the VM environment, the number of simultaneous processing workflows will increase.

This scalability is critical as the Moving Image Section looks to address the backlog in the born-digital moving image collections (non-AAPB), to create a sustainable, scalable processing model to ensure the best stewardship practice for the Library's collections and minimize or eliminate any processing backlogs.

VI. Conclusion

Performing analysis on the practical needs of born-digital projects and the impact of current policies that may not have born-digital workflows in mind are key to advocating the management of, if not more resources, then different approaches. Further, leveraging the expertise and experience of others working on similar projects creates a coalition when approaching management. Presenting a range of options increases the likelihood of finding a practical solution. Inevitably,

https://hdl.handle.net/2142/121129

the volume of born-digital projects the Library is encountering will only increase, and establishing foundations now, in documentation and adapting workflows to changing circumstances, will surely ensure future success in this endeavor.

Preserving born-digital moving image content presents many challenges and opportunities. Ever increasing file sizes and storage requirements, technical infrastructure, and maximizing processing throughout are a few of these challenges. By implementing a scalable processing and preservation program with IT support, these challenges can be minimized if not mitigated entirely. Automated workflows and digital file transfers versus manual processing and hard drive transfers are two actions that have significant impact in increasing productivity while being mindful of storage and technological infrastructure limitation. Such interventions demonstrate the possibilities that can arise with thorough thoughtful planning and the inclusion of additional resources.

References

[1] OpenCube
[2] BagIt. https://www.ietf.org/rfc/rfc8493.txt
[3] MySQL database. https://www.mysql.com/
[4] Baton Quality Control Software by interra Systems. https://www.interrasystems.com/file-based-qc.php
[5] FFmpeg. https://ffmpeg.org/
[6] MediaInfo. https://mediaarea.net/en/MediaInfo
[7] Dalet Amberfin. https://www.dalet.com/products/amberfin/