Seeing the big picture: Developing workflows for digital moving image content

Received (in revised form): 20th May, 2019

Rachel Curtis
is a digital project specialist at the Library of Congress and a project coordinator for the American Archive of Public Broadcasting (AAPB). In this capacity, she manages the ingestion of preservation files and associated metadata into the Library's archive, works with project partners on policy and strategy decisions and coordinates Library staff on AAPB activities. She holds a master's degree in library and information science from the University of Wisconsin-Milwaukee.

The Library of Congress, National Audiovisual Conservation Center, 19053 Mt. Pony Rd, Culpeper, VA 22701, USA
Tel: +1 202 707 3140; E-mail: rcur@loc.gov

Laura Drake Davis
is a digital project specialist in the Moving Image section of the Library of Congress. In this role, she processes born-digital moving image content, develops new workflows for born-digital content and develops strategies for metadata capture and transformation. She brings a wide range of experience to this role, with previous positions in college and university archives, special collections and state government archives. A certified archivist since 2007, she holds a master of library science degree from the University of Maryland College Park.

The Library of Congress, National Audiovisual Conservation Center, 19053 Mt. Pony Rd, Culpeper, VA 22701, USA
Tel: +1 202 707 0248; E-mail: ladavis@loc.gov

Abstract Workflow development is a critical aspect of successful project management. While time-consuming and documentation-heavy, project management is the key to the successful implementation of complex projects. This paper discusses workflow development at the Motion Picture Broadcasting and Recorded Sound Division of the Library of Congress, with a specific focus on moving image content. Sharing the evolution of digital processing and workflow development for moving image content, this paper discusses the efforts of the American Archive of Public Broadcasting (AAPB) and the establishment of positions dedicated to the processing of born-digital moving image content. The evolution of processes from early large-scale projects of the AAPB to the development of a fully automated workflow for the US Senate Floor Recordings are described with concepts applicable to organisations of any size and available resources.

KEYWORDS: digital collection, workflows, processing, project management, moving image collections

INTRODUCTION

Over the past few years, the Library of Congress (the Library) has received an ever-increasing amount of digital files, both born-digital and as the result of digitisation projects. While this is true for all the different divisions of the Library, from manuscripts to prints and photographs, the Library has only recently begun to identify and allocate dedicated resources to handling born-digital

audio-visual materials. The bulk of audio-visual material acquired by the Library is processed at the National Audiovisual Conservation Center, also known as the Packard Campus, which 'develops, preserves and provides broad access to a comprehensive and valued collection of the world's audio-visual heritage for the benefit of Congress and the nation's citizens'.[1] The Packard Campus is a stand-alone facility located about 70 miles south of DC in Culpeper, Virginia. As a result of this long-distance separation from the Library's hub in DC, Packard Campus is largely self-sufficient, with many services developed and provided in-house.

The Packard Campus is home to the Motion Picture, Broadcasting and Recorded Sound Division (MBRS). Inter-organisational collaboration within MBRS is strong and key to building and improving processing workflows. The bulk of material is received through more traditional acquisition paths, such as gifts and copyright deposit, but MBRS also supports community-based projects, like the American Archive of Public Broadcasting (AAPB) and the Silent Film Project. This paper is divided into two sections: the workflows developed for the AAPB and those developed for born-digital gifts and deposits.

PROCESSING BORN-DIGITAL AUDIO-VISUAL FILES BEFORE DIGITAL PROJECT SPECIALISTS

At the Packard Campus, workflows traditionally focused on processing and describing physical materials and ingesting the corresponding digitised files. The Packard Campus Workflow Application (PCWA), the Library's audio-visual ingestion software, was built around this model and operates primarily on 'ordered ingest'. This means that all files ingested into the system are originated as 'orders' based on the physical items held in the Library's collection. PCWA is closely interconnected with the Merged Audio Visual Information System (MAVIS), the collection management system used by MBRS. When these software configurations were implemented, file ingest was dependent on physical holdings, as the majority of the preservation work at MBRS was based on creating digital and access copies of the analogue collection. Most of the materials held by MBRS were physical tapes and reels, and born-digital, file-based material was not a large part of the collection. This changed over the past few years as the Library began to receive an increasing amount of born-digital audio-visual collections through various acquisition paths.

Ingesting born-digital collections requires an 'orderless ingest' functionality, where items without a corresponding physical item can be ingested. While PCWA supports this, it requires a different set of processes to operate and the functionality was neither well documented nor widely used. Born-digital material is acquired by the Library through various acquisition paths, including gifts and copyright deposit, and received via hard drive, LTO tape and digital deposit. Historically, these collections were not handled in a systematic fashion, but on an as-needed basis.

Hard drives were set aside, sometimes catalogued as items and shelved for later processing. A server space called 'embargo' was set up to store files received through digital deposit or offloaded from hard drive and LTO tape. This space is essentially a digital closet, where files can be stored, but are not easily accessible. Files on hard drive or in the embargo space were processed and ingested on an as-needed basis with no coordination between staff. This resulted in multiple workflows and duplicated, undocumented processes.

Over time, a large digital backlog began to develop in both the recorded sound and moving image sections. As more batches of digital files arrived, management recognised that current workflows were unsustainable and dedicated staff positions were needed to

process these files in a timely manner. This became especially critical when MBRS took on two new projects: History Makers and the AAPB.

THE AAPB

The AAPB began as a project funded by the Corporation of Public Broadcasting (CPB). CPB conducted an inventory project and then provided funds for 100 public television and radio stations to digitise items in their collection. This resulted in the creation of over 40,000 hours of digitised content and about 73,000 files. In 2012, CPB began looking for stewards to take over the AAPB project. The intent was to ensure the preservation of and accessibility to public media and find institutions with the capability to grow the collection beyond the initial 40,000 hours. CPB selected the Library of Congress and WGBH (a public broadcasting station based in Boston, MA) to be the co-stewards of the archive in 2013. The Library operates as the preservation arm of the AAPB, ingesting high-resolution preservation files into the archive and ensuring they are preserved for generations to come, while WGBH handles the access files, making them accessible through streaming on the AAPB website.

In 2013, the Library and WGBH received a grant from CPB, allowing the Library to hire a limited-term digital project specialist focused entirely on the AAPB. The vendor was scheduled to deliver the 70,000 files from the initial digitisation project in mid-to-late 2015 and there were several steps that needed to take place before these files could be ingested. Fortunately, the digital project specialist did not have to start from scratch. Many of the workflows used for file ingest were adapted from a recent acquisition of born-digital files from the History Makers project.

History Makers was the first major collection of born-digital material received by the Library that required immediate processing. As stated earlier, the ingest environment at the Packard Campus is geared towards digitisation, so this project became a test bed for the new workflows that had to be put in place to ingest born-digital material with no analogue carrier in the Library's collection. As there was no one on staff assigned to this work, development fell to the video lab supervisor, who was familiar with the technical specifications of PCWA and had the expertise to develop automated 'orderless' ingest workflows. Developing this new workflow involved reviewing PCWA's specifications, testing the orderless ingest capability, verifying successful ingest and automating as many processes as possible. These automated workflows became the template for those used for the AAPB, allowing staff to formalise these processes through documentation and implementing improvements.

The digital project specialist developed metadata mappings from CPB's archival management system to MAVIS, created documentation, coordinated ingest of the files and worked with the vendor to reconcile issues. As work on the AAPB moved forward and new grants were awarded to digitise more material, it was quickly apparent that a permanent position was required.

In 2015, the Library hired a permanent digital project specialist specifically for the AAPB project. This position would complete the initial ingestion of 70,000 files and coordinate several grant-funded projects recently awarded to the AAPB and contributors to AAPB, including:

- *PBS NewsHour Digitisation Project:* funded by the Council on Library Information Resources (CLIR), this project allowed a vendor to digitise over 8,000 episodes of PBS NewsHour's predecessor programmes from 1975 to 2007 (this project will be discussed in due course);
- *American Masters Interviews Digitisation Project:* funded by the National

Endowment for the Arts and undertaken by WNET (a public broadcasting station based in New York, NY), this project supported the digitisation of 800 raw interviews recorded between 1993 and 2012 for the award-winning PBS biography series 'American Masters' and
- *National Educational Television (NET) Cataloguing Project:* funded by CLIR, this project allowed the Library to hire two project cataloguers to process and catalogue the Library's extensive collection of NET material on film and tape.

In addition to these three projects, the AAPB team was reaching out to stations and producers to acquire their content. Much of that material is born-digital and again will be discussed in course.

The AAPB digital project specialist faced several challenges specific to the project:

- managing material from multiple sources in a variety of file formats;
- managing large, grant-funded projects involving up to four stakeholders;
- managing receipt of born-digital files from individual donors and
- making the case for open source tools (this will be discussed in due course).

Over 100 stations and producing organisations have participated in the AAPB, delivering over 80,000 preservation files to the Library in a variety of moving image and audio formats. Additionally, each institution has its own metadata and file-naming standards. The AAPB acquires material in two general categories: material on analogue tape that needs to be digitised and born-digital content that is already in a file-based format. For the former, institutions will apply for a grant and, if awarded, the Library will receive the preservation files directly from the vendor. AAPB staff work to guide the applicant institution through the process and coordinate project activities between all stakeholders (vendor, donor institution(s), the Library and WGBH). Material already in a file-based format is transferred to hard drive and delivered directly to WGBH and the Library. While material from the vendor arrives in the Library's preferred preservation format as specified in the contact with the vendor, born-digital materials arrive in their native formats. As a result, these files require more attention as they do not fit easily into automated ingestion workflows. The backbone of all the workflows relies on open source tools, such as OpenRefine, Media Info, FFmpeg, VLC Media Player and others, as these tools are harnessed to run checksums, harvest technical metadata, create access files and troubleshoot problems.

The following two case studies exemplify two different acquisition paths and their unique challenges.

CASE STUDY: THE PBS NEWSHOUR DIGITISATION PROJECT

In 2015, the CLIR awarded the AAPB with a Hidden Collections grant to digitise all the NewsHour predecessor programmes from 1975 to 2007. These programmes were held on a variety of deteriorating and obsolete tape formats. As a result of the grant, over 8,000 tapes were digitised over the course of two years.

This illustrates one way material comes to the AAPB — via a grant to support the digitisation of analogue material. This acquisition path is time-intensive, however; before the digitisation can begin, there is the long process of making a successful grant application and the logistical challenge of coordinating the delivery and receipt of analogue tapes and digital files between multiple stakeholders. All this work pays off, although, as the standardised file formats and monthly file delivery make it easier to implement automated workflows. Figure 1 provides an overview of this workflow.

There are two aspects to successful ingestion. As explained earlier, the AAPB leverages PCWA's 'orderless' ingest

quality of the source material. The QC process was added into the automated workflow, with files that failed needing manual review. The way files were received was also adjusted. Instead of getting files delivered on LTO tape, the Library requested hard drives instead. This facilitated offloading as it did not require access to the LTO tape robot, which is often being used other projects.

Files were delivered according to the Library's BagIt specification. BagIt is a hierarchical file system designed to support disk-based storage transfer of digital content. A 'bag' consists of a 'payload' (the content) and 'tags', which are metadata files intended to document the storage and transfer of the bag. A required tag file contains a manifest listing every file in the payload, together with its corresponding checksum. For the purposes of this project, the vendor was required to create a bag for each digitised asset. A workflow was created where bags were offloaded into a watch folder and the script would unbag them and run them through the entire process from initial validation checks, QC, ingest package creation and final ingestion into the archive. The only manual points were metadata creation, review of files that failed validation or QC and troubleshooting. Troubleshooting was also manually intensive, as the reporting features available through PCWA are not robust and sometimes it was not immediately apparent why a file failed the initial bag validation, in cases where the checksums matched.

Despite a few glitches, the system worked well, and the workflows developed were highly replicable because of the standardised nature of grant-funded projects. There was close cooperation with the vendor to ensure files were delivered in the required structure. These workflows have been adapted for other processes at the Library that require 'orderless' ingest.

CASE STUDY: PROCESSING BORN-DIGITAL COLLECTIONS

This case study will focus on AAPB collections donated by individual stations and producers. The workflow for acquiring and ingesting these collections is very different from material received through grant-funded programmes. Materials received this way are always born-digital or are the result of in-house digitisation. In some cases, this process is easier (and cheaper) as all that remains to do after an agreement is reached is to send the donor a hard drive and they copy their files onto it. This also means that there are a greater variety of file types, file names and directory structures. As a result, processing these files can be time-consuming because they do not fit into the Library's automated workflows.

A lot of work goes into identifying potential content for the AAPB. While a few stations reach out to the Library, the majority of content received this way is through outreach. The AAPB targets stations and producers with content that is under-represented in the archive. This does not always result in acquiring content, as many stations and producers have tight budgets and often lack the necessary resources to manage their archives. Even if a station cannot donate material, the conversation with them is always valuable for both parties, as it is possible to get an idea of what type of material they have, where they are with their own digital preservation plans and offer our expertise.

When reaching out to interested parties, the Library asks a few questions about their collection:

- What material do they have? We explain what material can and cannot be archived. Generally, the Library concentrates on full episodes of local programming or the raw interviews used in final productions.
- Do they own the rights to their material? Material made available in the AAPB Online Reading Room must clear a few copyright hurdles and this can be especially thorny with music programmes.
- Are their materials in analogue and/or digital format? Many archives have a mix

EXPANDING DIGITAL PROCESSING BEYOND AAPB

The work on the AAPB is just one part of the work performed in the moving image section. The work with the AAPB emphasised the need for a position devoted to born-digital content, and a new digital project specialist position was created in 2016 and filled in 2017. Responsible for the processing of gift and copyright collections, the digital project specialist also supports the creation and maintenance of websites containing moving image materials, including the National Screening Room and selections from the National Film Registry, maintains persistent identifiers for the moving image section and manages the creation of records for Library of Congress holdings in the Entertainment Identifier Registry (EIDR).

The Library had been receiving born-digital materials for several years, and without a dedicated position to process these materials, these materials became part of the backlog. These born-digital materials include a wide range of materials, including digital cinema packages, television programmes, recordings of sporting events, recordings from the US Congress, in-house restorations, videogames and other content. These collection materials were transferred via external hard drives and direct digital transfer. For materials on hard drives, the priority is to migrate the content off the hard drives onto network storage. An inventory of files revealed the variety of file types received — 117,833 files across 21 different formats. These files include 109,905 files in eight moving image formats.

The role of the digital project specialist is to process this material, with the goal of automating as many processes as possible. The automation portion of the goal is critical due to the volume of material held by the Library and the regular accruals of specific titles being received. For example, when the US Senate is in session, the Library receives floor recordings on a daily basis. An automated process allows the Library to maximise resources to create the descriptive record, create access files and ingest this content with minimal effort. More detail about this project follows in the next section.

An essential part of any of these projects is the project plan and the accompanying documentation. All of these projects have a project charter, project plan and a work breakdown. These documents are essential in documenting actions taken on these born-digital files and address the specific needs for each project.

The Project Charter consists of the following elements:

- *Project administrative information:* information regarding the project including name of project, responsible party/parties, accession/project identifiers and project start date.
- *Project background:* background information about the project including origination of project, project goals and impact on future projects, including development of processes that may be adapted to subsequent projects.
- *Project details:* specifics such as contents of collection by file and content type, accompanying metadata and project deliverables.
- *Project requirements (goals):* a description of the elements that will make the project successful, including project actions and deliverables.
- *Project risks:* the risks associated with not completing the project in a timely manner and not completing the project at all.
- *Approvals:* signatures of supervision to indicate approval of project plan and authority to begin project.

The project plan includes:

- *Project administrative information:* information about the project including accession number(s) and other unique identifiers, project owner(s), start date and proposed completion date.

- *Project scope:* the goals and outcomes of the project, to include those activities within the scope of the project and those that are out of scope.
- *Project assets:* the extent of the collection associated with the project.
- *Project deliverables:* the outcomes and deliverables of the project.
- *Stakeholders:* a list of the key stakeholders along with their role and responsibilities related to the project.
- *Timeline:* project milestones, projected and actual start dates and projected and actual completion dates.
- *Resource requirements:* the resources required to complete the project, regardless of administrative area of the institution.
- *Communications plan:* an outline of the plan for communicating with stakeholders, to include type of communication, how the communication is to be delivered, frequency of communication and the communication owner.
- *Document history:* a history of the project plan document, including creation and revision history, author and date of action.
- *Document specifications:* information regarding where the project plan is located, including document name and location.
- *Related documents:* a list of related documents associated with the project (project charter, work breakdown, inventories, etc).

The work breakdown is the log of activities performed on the project. This serves as documentation of the progress on, processes utilised and obstacles encountered on the project. While maintaining the work breakdown can be tedious, it is useful to document successes, failures, delays and other elements of a project. This information can be used to inform process improvement and communicate to management review resource allocation.

When working with born-digital content, the aim is to develop efficient processes for each project, recognising the specific digital assets for the project and the desired outcomes. Here, it is essential to identify the correct tools to utilise for the project itself — not just tools already in use at the institution, but also new tools that can be incorporated into this and future projects. The tools used to process collections include Python, MediaInfo and FFmpeg. The following case studies illustrate three projects and the approach to addressing the needs and challenges of each project.

CASE STUDY: US SENATE FLOOR RECORDINGS

The first project to utilise a fully automated workflow is the US Senate Floor Recordings project. The Senate floor recordings are produced by the Senate Recording Studio and document activities in the Senate Chamber. The Senate delivers floor recording files as 1-hour blocks, with each block containing a moving image file, an XML metadata file and an XML closed-caption file. Each legislative day can last anywhere from less than 1 hour to over 24 hours. Thus, the number of files received each day can vary considerably. For example, a 13-hour legislative day will result in 39 files transferred to the Library of Congress (13 MXF, 13 XML and 13 closed-caption files).

This project is complex as there are four groups of content that need to be processed, depending on when the files are transferred. The four groups are: (1) current, daily receipts; (2) files received prior to May 2018 (the December 2015 to April 2018 backlog); (3) files yet to be received (2007 to December 2015); and (4) closed-caption files received for materials already ingested (December 2015 to October 2018), that were not available at the time of initial transfer.

In reviewing the goal of building an automated process and the four groups of materials, a strategy was developed to build the initial workflow using the backlog files and then build in the automation to address

- *Posted to website:*
  - files delivered;
  - website updated.

For the National Screening Room, films are selected and scanned. As the film scanning process scans the film from edge to edge, the digital files are edited to the exterior of the film frame and speed-corrected as needed (early film were produced at 16 frames per second; film produced since 1927 is generally 24 frames per second). Once the files have been edited, the file is deposited in a directory where the files for the website are created with the Library of Congress bumpers at the beginning and end of the title. Following this, the thumbnail images are created for use on the website. These thumbnails are created to reflect the essence of the film or reflect a well-known scene from the title. WA Python script was developed using FFmpeg and MediaInfo to capture JPG images for the media player background images. This script relies on the user entering a fractional point within the file from which to extract a still image to be used as the thumbnail. That value is entered into the script, and the program generates thumbnail images for all of the files within the directory. Sometimes it takes several attempts to get representative images. For titles with a very short running time, it is sometimes necessary to view the title as the bumpers may be as long as or longer than the film itself. The best images are selected and submitted for approval. For the approved images, IrfanView is used to create the GIF images for the title display on the website.

Metadata in the National Screening Room is generated from one of two sources — the Library of Congress catalogue or from MAVIS. Retrieval of metadata from the Library of Congress catalogue for use in the National Screening Room is accomplished through an established process within the MARC record. For those records where these is no Library of Congress catalogue record, a Python script is used to pull the metadata from the MAVIS record and export it to a locally developed tool created specifically for item-level metadata not usually found in a library catalogue.

When files are submitted for inclusion on the website, another Python script is used to create the media ingest document. This document is used to connect the digital content with the metadata source for the presentation on the website. The information pulled for this document includes the local identifier for the content and the Library of Congress Control Number (if the item is in the Library of Congress catalogue).

Various tools are used to generate the products necessary to prepare and submit the

Table 2: Software used for the National Screening Room project

Software used	Content selected	Moving image content gathered	Still image files generated	Metadata review and creation	Files and metadata submitted	Content posted to website
Python		✓	✓	✓		
PyCharm			✓	✓		
FFmpeg		✓	✓			
VLC Media Player			✓			
Media Info		✓	✓			
Trello	✓	✓	✓	✓	✓	✓
Confluence	✓			✓	✓	✓
IrfanView			✓			

files and metadata for inclusion in the National Screening Room (see Table 2). Tracking the various titles and the status of any one title within the multi-step process is challenging, and progress is tracked with a Kanban (Trello) board and a series of spreadsheets.

Unlike the US Senate floor recording project, the scripts for this project have limited reuse potential, but have been repurposed for other website projects, including the Freud Home Movies and the Geographers on Film Project.

In this case study, consultation with partners across the Library was critical to ensure that the workflows were developed in accordance with existing procedures in other areas. Using that knowledge, it was possible to look for ways to automate some of the processes and documentation to increase efficiency.

CONCLUSION

These four case studies illustrate the iterative process of born-digital processing. Each process has built off the ones that came before, borrowing elements and improving them. These then inform older processes, which are brought up to date as new procedures are put in place. For example, the workflow laid out in the NewsHour case study is used for all vendor-submitted files, with the process tweaked for each new collection, based on lessons learned from the other workflows. The workflow for the US Senate Floor Recordings will serve as a model for the creation of other fully automated workflows. They illustrate the dynamic relationship and close collaboration between the digital project specialists and between them and the Library at large.

Having access to the right tools to process files and troubleshoot problems is a must. The digital project specialists at the Library spend part of their time identifying and analysing available tools adopted by the digital preservation and archiving community to improve processing and create efficiencies. For security reasons, the Library requires software new to the Library be thoroughly evaluated before adoption. The digital project specialists make the case for tools to be officially supported by the Library and work with other departments in the Library that perform similar tasks to coordinate requests.

Collaboration, not only at the interdepartmental level, but with the archival community at large is vital to the success of any digital preservation programme. Within MBRS, there are individuals who have programming skills, deep knowledge of the software systems and familiarity with Library policies. Library staff in Washington, DC, engaged in similar endeavours provide assistance and added perspective to Library activities, easing collaborative opportunities despite the physical separation of the Packard Campus. At the community level, the Library engages with organisations such as the Association of Moving Image Archivists (AMIA), the American Library Association (ALA), the Society of American Archivists (SAA) and the Digital Library Federation (DLF) and discusses its processes, shares its workflows and listens to others explain theirs.

The development of workflows on a project-by-project continues, but is informed by the work described within this paper. The Library continues to look for new tools and process improvements. Each project must be evaluated to assess the resources available and strategies developed to ensure outcomes consistent with other projects. While full automation is not always possible for every project, great strides are being made towards to this goal, and staff are better informed about what is needed to enable efficient born-digital processing workflows.

References

1. Library of Congress. (n.d.) 'The Packard Campus: Mission', available at: https://www.loc.gov/avconservation/packard/mission.html (accessed 4th April, 2019).
2. Library of Congress. (n.d.) 'National Screening Room: About this Collection', available at: https://www.loc.gov/collections/national-screening-room/about-this-collection/ (accessed 23rd March, 2019).