Federal Agencies Digitization Guidelines Initiative

Creating and Archiving Born Digital Video

Part II. Eight Federal Case Histories

September 8, 2014

The FADGI Audio-Visual Working Group

http://www.digitizationguidelines.gov/audio-visual/

Creating and Archiving Born Digital Video II: Eight Federal Case Histories

By the Federal Agencies Digitization Guidelines Initiative Audio-Visual Working Group
http://www.digitizationguidelines.gov/audio-visual/

Version 1.0, September 8, 2014

TABLE OF CONTENTS

INTRODUCTION

WHAT IS THIS DOCUMENT?

This is one of four documents examining aspects of the current practice for creating and archiving born digital video at selected institutional members of the Federal Agencies Digitization Guidelines Initiative Audio-Visual Working Group. The three companion documents are:

- *Creating and Archiving Born Digital Video I:Introduction (Version 1.0, 9/8/14)*
- *Creating and Archiving Born Digital Video III: High Level Recommended Practices (Version 1.0, 9/8/14)*
- *Creating and Archiving Born Digital Video IV: Resource Guide (Version 1.0, 9/8/14)*[1]

UNDERSTANDING THE CASE HISTORIES

As discussed in greater detail in *Creating and Archiving Born Digital Video I: Introduction*, the diverse set of eight case histories included in this report explore aspects of the current state of practice in six federal agencies working with born digital video. Each case history details the project deliverables and technical specifications but also strives to tell the story of the project, including the background of the institution and the collection but also the goals and lessons learned. The projects covered in the case histories are examples of one of many projects within larger institutions which may have different requirements, outputs and workflows.

The case histories intersect with the Recommended Practices outlined in *Creating and Archiving Born Digital Video III: High Level Recommended Practices* which emerged from the collective project experiences.

The eight case histories are organized into two groups: *Creating Born Digital Video* and *Archiving Born Digital Video*. Within each group, case histories are listed in alphabetical order by contributing institution.

Creating Born Digital Video Case Histories

- **LC-AFC-CRHP:** Library of Congress American Folklife Center Civil Rights History Project
- **NOAA-OkEx:** National Oceanic and Atmospheric Administration Okeanus Explorer
- **VOA-MMAM**: Voice of America Metadata for Media Asset Management

Archiving Born Digital Video Case Histories

- **LC-NAVCC-VEF:** Library of Congress Packard Campus of the National Audio-Visual Conservation Center Video Evergreen Format
- **LC-WebArch-YouTube:** Library of Congress Web Archiving You Tube Harvesting
- **NARA-BRCC:** National Archives and Records Administration Base Realignment and Closure Commissions project
- **SIA-DVD**: Smithsonian Institution Archives Authored DVD project
- **SI-DAMS:** Smithsonian Institution Digital Asset Management System

[1] The URLs for the three documents are:
(I) http://www.digitizationguidelines.gov/guidelines/FADGI_BDV_p1_20140908.pdf
(III) http://www.digitizationguidelines.gov/guidelines/FADGI_BDV_p3_20140908.pdf
(IV) http://www.digitizationguidelines.gov/guidelines/FADGI_BDV_p4_20140908.pdf

The goal of the three *Creating* case histories is to encourage a thoughtful approach from the very beginning of the video production project, before even shooting the video, which takes sustainability and interoperability into account. The five *Archiving* case histories tell the story of bringing the born digital video files into managed data repositories for long term retention and access and explore the issues which emerge when the born digital video objects arrive at the archive.

The last section of this document contains summary tables of the file specifications implemented in the case history projects.

CREATING BORN DIGITAL VIDEO CASE HISTORIES

Three of the case histories include the creation of new born digital video files as part of the project. These projects actually extend beyond just the creation cycle because these files are eventually ingested into managed repositories. The goal of the *Creating* case histories is to encourage a thoughtful approach from the very beginning of the video production project, *before* shooting the video, which takes sustainability and interoperability into account because choices made during the file creation process will have impacts on the long term archiving and distribution processes. They reflect Recommended Practices which illustrate the advantages of starting with high quality data capture from the very start.

Each case history focuses on a different problem set. LC-AFC-CRHP describes the challenges of creating a large collection in a short timeline. NOAA-OkEx describes how space limitations on board ship required adjustments to the capture and storage workflows. VOA-MMAM describes the metadata workflows within their in-house systems.

LIBRARY OF CONGRESS
AMERICAN FOLKLIFE CENTER
CIVIL RIGHTS HISTORY PROJECT:
FROM FIELD TO REPOSITORY TO PORTAL

CASE HISTORY SUMMARY

To fulfill the congressionally mandated Civil Rights History Project Act of 2009,[2] the American Folklife Center at the Library of Congress and the Smithsonian's National Museum of African American History and Culture jointly undertook a video documentation project beginning in 2010. The goal of the initiative was to record the experiences and memories of participants in the struggle to secure civil freedoms for African Americans in the US, broadly covering events and actions from the 1950's to the 1970's. The project's five year time frame leaves little room for error or delay. Within this period, AFC and NMAAHC are to capture dozens of interviews and very shortly thereafter provide public, research access to the interviews, including the preparation of EAD finding aids, complete bibliographic records, and the production of video streams of the recordings, primarily through the LC's and SI's web portals. These challenges necessitated technical solutions that are unlike those previously employed for similar large-scale oral history and ethnographic documentary productions at the AFC. The archive encompasses millions of items of ethnographic and historical documentation recorded from the nineteenth century to the present. Included are nearly 400,000 hours of recordings, more than a half-million photographs, several million pages of manuscript materials, and artifacts.

CASE HISTORY AUTHORS

Guha Shankar (gshankar@loc.gov), American Folklife Center, Library of Congress

Bert Lyons (blyo@loc.gov), American Folklife Center, Library of Congress

Carl Fleischhauer (cfle@loc.gov), Office of Strategic Initiatives, Library of Congress

John Bishop (john@media-generation.com) Media Generation (contractor)

INTRODUCTORY INFORMATION

Institutional Background

The American Folklife Center[3] (AFC) at the Library of Congress is the premier repository for ethnographic audio-visual materials in the world. Its holdings include hundreds of hours of unique, historic collections of spoken word, verbal art and oral performances, including the recorded testimonies of formerly enslaved African Americans and the first-ever field recordings -- the 1890 cylinder recordings of the anthropologist J.W. Fewkes of Passamaquody Indians from Maine, along with similarly rare items. The archive encompasses millions of items of ethnographic and historical documentation. Included are nearly 400,000 hours of audio-visual recordings, more than a half-million photographs, several million pages of manuscript materials, and artifacts.

AFC has a staff of approximately 23 FTE. The principal staff for the Civil Rights History Project[4] (CRHP) includes a Project Director to coordinate the LC's efforts with its SI partners, along with a dedicated, full-time Processing Archivist/Cataloger. These two individuals work in conjunction with the AFC Digital Assets Manager to process and prepare the collections for long-term, archival deposit and for public access. Additional personnel consist of a two-person interview team—a subject specialist on the topic of Civil Rights who is also the lead interviewer, and a professional filmmaker/camera operator. These staff members are hired by the field production coordinator, UNC

[2] http://www.gpo.gov/fdsys/pkg/PLAW-111publ19/pdf/PLAW-111publ19.pdf
[3] http://www.loc.gov/folklife
[4] http://www.loc.gov/folklife/civilrights/

Chapel Hill's Southern Oral History Program[5], under contract to the National Museum of African American History and Culture[6] (NMAAHC).

Collection Background

The Civil Rights History Project Collection[7] (AFC2010/039) contains, principally, the video documentation of oral history interviews with a wide spectrum of participants or "veterans" of the Civil Rights Movement (ca. 1954-1978) across the nation. These are born digital (see exception below) videos, captured in high definition format. As of summer 2014, one hundred six interviews are available on the LC's public search interface[8] and through iTunes.[9]

The project met the challenge to produce high-quality, sustainable video files by engaging the services of a professional filmmaker with high-quality field recording equipment and knowledge of the challenges of shooting for both production and preservation.

CASE HISTORY DETAILS

Within the five year project timeline, AFC and NMAAHC are to capture hundreds of interviews and shortly thereafter provide public access for researchers and the general public to the interviews and associated metadata primarily through the LC's and SI's web portals.

Specific project research areas include:

- Establishing video production and capture specifications
- Creating EAD finding aids
- Creating complete bibliographic records
- Digital file management
- Establishing flexible and scalable workflows

Selecting the Right Equipment

Typically, a high quality video camera generates an uncompressed video stream with 4 bits of luminance information, 4 of red color, and 4 of blue color notated as 4:4:4 (which has a bandwidth of 1.485 Gb/sec). The camera then down-samples this to 4:2:2 for high-end cameras or 4:1:1 for most consumer ones. The video frames are compressed into groups of 16 pictures (GOPs) which contain one full resolution frame followed by 15 frames extrapolated from changes calculated from the original frame. In effect, only two full frames per second are stored and the other 28 are reconstructed by the playback device. This is stored on a solid-state storage device, typically an SD card, at less than 25mb/sec bandwidth.

For the CRHP, the equipment to record an uncompressed 30 fps 1080 4:4:4 signal was too expensive, too cumbersome for shooting on location in almost as many cities as there were informants, and required specialized storage that was also prohibitively expensive.

To achieve the project goals and stay within budget, the filmmaker's choice of camera was the Sony XDCam EX-1 camera. It has three chips which keeps the RGB images discrete, instead of being extrapolated from a Bayer filter. The ½" chips are larger than the ¼" or ⅓" chips in most other cameras, allowing the operator to shoot with shallower depth of field, thereby minimizing the focus of the background. It accepts XLR balanced microphone inputs and supplies 48v phantom power to the microphones and it has level controls. It has an HD-SDI output which provides both audio and video to the AJA KiPro (see next paragraph). The operator can adjust focus, framing, exposure, and sound levels on the camera. A small LCD video monitor and headphones can be fed from the AJA KiPro as a confidence check that all elements of the recording chain are working.

---

[5] http://www2.lib.unc.edu/dc/sohp/sohp.html
[6] http://nmaahc.si.edu/
[7] http://www.loc.gov/folklife/civilrights/
[8] http://www.loc.gov/collection/civil-rights-history-project/about-this-collection/
[9] https://itunes.apple.com/us/itunes-u/civil-rights-history-project/id880114010

The AJA KiPro is a device that transcodes most HD signals into the Apple ProRes 422 (HQ) codec and records it on a removable hard drive. ProRes is an intermediary codec designed for original recording that will hold up well through color correction, titling, and compositing. This codec retains 4:2:2 information and compresses each frame individually (intraframe) as compared to the GOP (interframe) approach. It records a bandwidth of 220 mb/sec. In practice, we recorded the uncompressed 4:2:2 30 fps HD stream from the SDI output of the camera so as to record a little more than two hours on a single 250 GB replaceable KiPro drive, which could be swapped out in less than a minute for a blank drive.

Opting for Lossy Compression

This project decided to use compression because we were not convinced that uncompressed video or even higher resolutions (e.g., 2k or 4k) was necessary for capture, given that the oral history interview format is generally of a different order of complexity and dynamism than a Hollywood style feature length movie. The latter genre demands extremely high production values, in anticipation of complex editing and compositing tasks in the post-production phase.

Files are encoded utilizing Apple's ProRes HQ codec (also known as Apple ProRes 422), a lossy codec. It is an intermediary codec designed for original recording that will hold up well through color correction, titling, and compositing. This codec retains 4:2:2 information and compresses each frame individually (intraframe) as compared to the GOP (interframe) approach. It records a bandwidth of 220 Mbps.

SUMMARY OF VIDEO AND AUDIO DATA TECHNICAL SPECIFICATIONS

Video Data

- Container/wrapper: QuickTime (.mov); MPEG-4 format
- Target total bitrate: 220 Mbps (variable bit rate mode)
- Timecode: SMPTE
- Frame size: 1920 x 1080
- Aspect ratio: 16:9
- Video codec: ProRes HQ (422)
- Compression type: Lossy
- Video frame rate: 29.97 fps
- Color encoding: YCbCr
- Chroma format: 4:2:2
- Bit rate: 10-bit
- Interlaced scan

Audio Data

- Audio channels: 2
- Audio codec: PCM
- Audio sample rate: 48 kHz
- Audio bit depth: 24 bit

RELEVANT RECOMMENDED PRACTICES

This project is a great example of the "make the best file you can afford to make and maintain" ethos. The project goals permitted following a variety of the Recommended Practices including:

Advice for File Creators

- **Select a camera and other recording equipment with capability to capture at high quality levels:** All capture decisions are a balance of technical capacity in the digital capture equipment and the project goals for quality levels. It is best to start with a camera or capture device such as the Sony XDCam EX-1 used for this project that errs on the side of technical complexity, rather than be limited in capture specifications due to a device that is too basic for the job.

- **Provide the means to collect and submit metadata starting at the video shoot:** This project provided the interview team and production coordinator with the means to submit metadata to repository in real time, or as soon as the interview event is concluded. The collaborative cataloging application built for this project lives at https://lcapp.loc.gov/afccp (username and password required). Project team members have access to this website to create interview records, file-level records, and name authority records for the interviewers and interviewees.

- **Capture video data to stable storage devices that allow for easy file transfer:** The interviews are captured via the camera directly to an external hard disk (the AJA KiPro), at the point of recording. NOTE: This is the standard capture and delivery format and settings for all the interviews in the CRHP except for the first six interviews that were captured to HAEM videotape.

- **Select High Definition (HD) rather than Standard Definition (SD):** The recording is in the High-Definition format, 1080/60i.

- **Select larger picture sizes over smaller picture sizes:** Frame resolution is 1920 x 1080 pixels.

- **Select higher bit rates over lower bit rates:** LC-AFC-CRHP data rates almost uniformly achieve 220 Mbps for each segment or video file—an interview may be comprised of several files, due to the method employed by the shooter of "writing" a given amount of interview material to the hard disk to preclude losing the entire file if corruption occurs in any part of the interview.

- **Select higher bit depths over lower bit depths**: LC-AFC-CRHP captures in 10-bit.

- **Use higher chroma subsampling ratios rather than lower:** LC-AFC-CRHP captures in 4:2:2 which is supported by ProRes.

- **Generate a high integrity and continuous master timecode:** LC-AFC-CRHP creates SMPTE timecode.

- **Stay within the range of common frame rates of 24 - 30 frames per second (fps):** LC-AFC-CRHP uses 29.97 fps.

Advice for File Archivists

- **Identify the file characteristics at the most granular level possible, including the wrapper and video stream encoding**: Because LC-AFC-CRHP established the target format specifications for this project there are few surprises. AFC archivists use MediaInfo and Exiftool to verify that each file meets the expected profile.

- **Move video files off internal camera data storage, videotape, optical media or other unstable physical carriers to more stable storage media as soon as possible**: The most sensitive (precarious) aspect of making these recordings was making redundant copies of the files to separate drives, while in the field. The procedure was to write individual files to the KiPro (external hard disk recorder) in the process of capturing the interview, so that several files of 15–20 minutes in duration were discretely captured. At the end of each day's filming, the KiPro drives were attached to a computer via Firewire and copied to two hard-drives. Then the KiPro drives were erased for the next day's interviews. One backup drive was in hand luggage while traveling. Upon delivery to the archive, original files are bagged (using BagIt specs) and copied to managed servers for long-term storage.

Advice for File Creators and File Archivists

- **Select video encoding and wrapper formats that are well-supported now and future focused:** ProRes is established, well documented and widely used, particularly in the postproduction environment.

- **Select video encoding and wrapper formats that are non-proprietary:** LC-AFC-CRHP uses ProRes, a proprietary format, because it retains the 4:2:2 information and compresses each frame individually (intraframe) as compared to the GOP (interframe) approach. Files are wrapped in Quicktime .mov wrapper, a well-documented wrapper format.

- **Select video encoding and wrapper formats that are supported by downstream applications:** ProRes HQ codec and the Quicktime wrapper are fully compatible with NLE's using Final Cut Pro software presently installed in the suite and eliminates the need to transcode the content or render the video into a lower bit rate/resolution for editing and production.

- **Select video formats that are standardized and well documented:** ProRes is well documented.

- **Select formats that can contain and label complex audio configurations including multiple channels and sound fields beyond mono and stereo:** ProRes supports up to four audio channels.

When Following the Recommended Practices Is Not Practical

The project goals required some compromise and preclude following several Recommended Practices most notably in the choice of lossy compression:

- **Select uncompressed over compressed:** LC-AFC-CRHP implements compression because the relatively simple content, oral history interviews, did not warrant the larger file files from uncompressed video.
- **If compression is used, select mathematically lossless compression over visually lossless or lossy compression:** LC-AFC-CRHP used ProRes HQ (422) lossy compressed because the relatively simple content, oral history interviews, did not warrant the larger file files from uncompressed video.
- **Select open formats over proprietary formats:** LC-AFC-CRHP uses Pro-Res, a proprietary format, because it retains the 4:2:2 information and compresses each frame individually (intraframe) as compared to the GOP (interframe) approach.

LESSONS LEARNED

File Management and Storage

Even though this project selected a lossy compressed format, the large file sizes have been the biggest obstacle of the project. The delivery of enormous files ranging from a minimum of approximately 30 GB to upwards of 350 GB per interview severely taxes the local production and storage environment. The key constraints in preserving this collection have been a combination of time and institutional throughput.

The extent (size) of the files in this project must be understood in the context of inherent challenges in managing the AFC's extant digital collections which number over twenty-five hundred. AFC's collection are just some of the several thousand other such collections in the LC as a whole. The modest collection of 108 interviews in the CRHP consume18.33 TBs of space, at an average of 169.69 GBs per interview. But these numbers refer only to the size of the original master copies for each interview. There are additional storage costs such as the size and number of derivatives - backups, mezzanine/production copies, and service copies - which more than double the storage requirements of the masters. If we calculate the cost of this storage solely based on master files based on Amazon storage pricing ($0.03 per GB per month as of spring 2014), the estimated return is as follows:

Average Size of Interview	Cost Per GB Per Year	Cost of Storage Per Interview Per Year	Total Size of Original Collection	Cost of Storage of Collection Per Year
169.69 GB	$0.36	$61.08	18,330 GB (18.33 TB)	$6,599

Taking into account the complete set of files that make up the extent of this collection as noted above, the cost for storage of this one collection of born digital video could easily approach $12,000 per year.

Metadata Workflows

Design a shared cataloging utility that enables workers—interviewers, documentarians—and others in the field, that is, off-site, to provide descriptive and contextual metadata at the point of capturing footage. This is a corollary action that supports one of the best practices of ethnographic fieldwork and is captured by the phrase, "the fieldworker is the first archivist." This method significantly enhances processing and cataloging of the interviews, which are often undertaken by repository staff who are not centrally involved with the documentation process and are often completed at a point in time well after the interviews are received by the repository. The result has been that the first set of fifty-six discrete interviews (of the total 103 interviews) has been fully processed, cataloged, and made available for discovery and access online within two years after the project's launch.

Infrastructure Upgrades

AFC implemented several emerging (just-in-time) solutions:

- In order to produce viewing copies—on DVD for the partner institution and participants, along with a file-based mezzanine intermediate master format from which to produce streaming videos for the web-based portal—a dedicated digital editing work-station with multi-core processing capabilities must be utilized. This required a system upgrade from previous generations of Apple G5 units to Intel-based Macintosh computers with fairly significant amounts of RAM. We assume that this scaling up is not news to anyone involved in AV production, whatever platform they are working in.
- An external NAS, capable of storing several TBs of data is in place to accomplish editing and concatenation of the several files that comprise a single interview. This process is quite distinct from the one of ingesting the original master files to managed LC servers for long-term storage, but the local production environment must be reckoned with before embarking on a collecting project of this scale.

NATIONAL OCEANIC AND ATMOSPHERIC ADMINISTRATION
OFFICE OF EXPLORATION AND RESEARCH
OKEANOS EXPLORER DEEP SUBMERGENCE VIDEO PROGRAM

CASE HISTORY SUMMARY

NOAA's Office of Ocean Exploration and Research[10] Deep Submergence Group collects video data from a 6,000-meter-rated Remotely Operated Vehicle (ROV) system. The vehicle is operated from the NOAA Ship *Okeanos Explorer*,[11] "America's Ship for Ocean Exploration," - the only federally funded U.S. ship assigned to systematically explore our largely unknown ocean for the purpose of discovery and the advancement of knowledge.

A key program goal is rapid and broad data dissemination. During an ROV dive, several video capture and production activities are performed simultaneously to enable both immediate data sharing as well as to meet longer term data management goals. Using telepresence technology, video data collected at ROV operational depth are transmitted in real time from the ROV to the ship – then to shore – and streamed to the public and shore-side cruise participants. Simultaneously, shore-side cruise participants viewing the video streams collaborate with the shipboard scientists to describe and guide the dive, while annotating the observations via chat logs. On board the ship, videographers analyze the video data streams and capture key items of interest for long-term use. The streamed video is captured aboard the ship for post cruise analysis.

This complex set of activities results in a broad array of video data, data products and annotation logs. On-board management of this collection is complicated by: a) the limited physical space aboard ship for data storage devices; b) the video workflow itself, in which the production and management systems have differing requirements; and c) the limited transmission pipeline off the ship.

CASE HISTORY AUTHORS

Brendan Reser (brendan.reser@noaa.gov), General Dynamics Information Technology / National Oceanographic Data Center on assignment with the NOAA Office of Exploration and Research

Sharon Mesick (sharon.mesick@noaa.gov), National Coastal Data Development Center

INTRODUCTORY INFORMATION

Institutional Background

NOAA's Office of Ocean Exploration and Research (OER) leads partnerships to accomplish national ocean exploration goals; conducts interdisciplinary characterizations of unknown or poorly known ocean areas; improves the technical capabilities of the marine science community and engages and educates audiences in ocean exploration using innovative means.

NOAA Ship *Okeanos Explorer*, "America's Ship for Ocean Exploration," is the only federally funded U.S. ship assigned to systematically explore our largely unknown ocean for the purpose of discovery and the advancement of knowledge. Aboard the Okeanos, OER's Deep Submergence Group collects video data from a 6,000-meter-rated Remotely Operated Vehicle (ROV) system. Telepresence,[12] using real-time broadband satellite communications, connects the ship and its discoveries live with audiences ashore.

Via telepresence, up to 3 720p video streams are transmitted from the ship real time during ROV operations. Additionally, real time ship navigation, meteorological and in situ observations, as well as sonar data products are synchronized between shipboard and shore-side data repositories. This enables scientists on board ship and on shore

[10] http://explore.noaa.gov/
[11] http://explore.noaa.gov/Exploration/OkeanosExplorer.aspx
[12] http://oceanexplorer.noaa.gov/explorations/07blacksea/background/telepresence/telepresence.html

to access the same datasets in the same timeframe for joint analysis. During some sea-going expeditions, more than 100 shoreside participants have been engaged in the mission using telepresence technology.

Audio and written communication between shoreside participants and the shipboard crew is managed via RTS intercom units, teleconference via VoIP, or a chatlog. The diversity of the scientists engaged via telepresence-enabled data sharing ensures that subject matter experts are available to characterize the environment and contribute to operational direction as the ship and ROV systems explore, regardless of whether the expedition focus is exploration of biological, geological, or archaeological features.

A broad range of staff are engaged in these operations, including: expedition coordinators, ROV engineers and operators, videographers, telepresence engineers, and data managers. These individuals may participate during the planning, the at-sea collection, and/or following a cruise in the on-shore data curation. The current video collection is approximately 45TB in volume with an expected increase of 16 – 30 TB / year from FY2014 moving forward.

Collection Background

The *Okeanos Explorer* video collection is comprised primarily of subsurface video gathered using a two-vehicle deep submergence system. Two vehicles operate together with one functioning as a camera sled (ROV Seirios[13]) and one as the primary operational ROV. The purpose of the camera sled is to mitigate wave action propagating down the tether to the surface support platform (*Okeanos Explorer*), to provide positional awareness for the primary vehicle, and to provide focused high intensity lighting. From 2008 to 2012 the primary operating vehicle was the ROV Little Hercules[14]. In 2013 the OER program brought a new vehicle online, the ROV Deep Discoverer.[15]

Video is routed from these systems via fiber to the ship and into an EVS instant replay system. Clips are selected from the captured footage and exported as ProRes 422 within a QuickTime wrapper. Historically clips have been selected by shipboard videographers as well as participating scientists. As of FY14 the program intends to expand its workflow such that it will capture end to end (dive to ascent) video footage.

CASE HISTORY DETAILS

As a federally owned vessel outfitted and primarily used for exploration, the *Okeanos Explorer* is tasked with providing an initial foray into previously unexplored / researched areas. Target regions are determined based on input from both the scientific and policy driven community. Since its inception, the *Okeanos Explorer* has explored regions in Indonesia, the Hawaiian Islands, the North American Pacific shelf, the Galapagos spreading center, the Mid-Cayman rise, the Gulf of Mexico, and canyons and sea mounts along the North Atlantic shelf break. ROV dive targets have included such diverse topics as methane seeps, hydrothermal vents, coral communities, and marine archaeological sites. The nature of the telepresence paradigm ensures that scientific subject matter experts in all topic areas are available to review, identify and provide guidance on whatever discoveries may be made.

Video data collections provide a ‘do no harm’ approach to documenting the ocean environment. The unique operational methods on board the *Okeanos Explorer* generate a large quantity of original source data, video data products at various resolutions, and documentary footage of shipboard operations.

There is both a fiduciary and scientific responsibility to preserve these data for the long term. Determinations of what data to preserve, for how long and in what formats are key to this research.

[13] http://oceanexplorer.noaa.gov/okeanos/explorations/ex1103/background/seirios/seirios.html
[14] http://oceanexplorer.noaa.gov/okeanos/explorations/10index/background/rov/rov.html
[15] http://oceanexplorer.noaa.gov/okeanos/explorations/ex1302/new_rov_video.html

SUMMARY OF VIDEO AND AUDIO DATA TECHNICAL SPECIFICATIONS

Video Data

- Container/wrapper: Apple QuickTime (.mov)
- Target total bitrate: 145 Mbps
- Timecode: SMPTE (Control Clock set to UTC)
- Frame size: 1920 x 1080
- Aspect ratio: 16:9
- Video codec: ProRes (422)
- Compression type: Lossy
- Video frame rate: 29.97fps
- Color space / encoding: YUV
- Chroma format: 4:2:2
- Interlaced scan (Top Field First)
- Video frame rate: Constant
- Color Primaries / Transfer Characteristics / Matrix coefficients: SMPTE 240M

Audio Data

- Audio channels: 4
- Audio codec: PCM
- Audio sample rate: 48 kHz
- Audio bit rate: 1152 Kbps
- Audio bit rate mode: Constant
- Audio bit depth: 24 bit

RELEVANT RECOMMENDED PRACTICES

The project goals permitted following a wide variety of the Recommended Practices including:

Advice for File Creators

- **Select a camera and other recording equipment with capability to capture at high quality levels:** Due to the nature of the data collection the cameras that the *Okeanos Explorer* program uses in video capture are custom built specifically for the ROV's using them. In the case of the Seirios, the Little Hercules, and the Deep Discoverer all systems use either an Insight Pacific Zeus+ or a Zeus+ in combination with an Insight Pacific miniZeus. All are HD cameras supporting 1080i.

- **Provide the means to collect and submit metadata starting at the video shoot:** The OER data management team generates metadata at the end of each cruise.

- **Capture video data to stable storage devices that allow for easy file transfer:** Video is captured directly to spinning disk on an EVS instant replay system. Video is then clipped out and saved to a large Nexsan SAN array.

- **Select larger picture sizes over smaller picture sizes:** File resolution is 1920 x 1080 pixels.

- **Select higher bit rates over lower bit rates:** Video is captured at 145 Mbps. While this is not the highest available bit rate available via ProRes (specifically not the highest available using ProRes HQ), nor is it the lowest. 145 Mbps meets the need for gathering the highest quality video in limited storage.

- **Use higher chroma subsampling ratios rather than lower**: The *Okeanos* video team captures in 4:2:2, which is supported by ProRes

- **Generate a high integrity and continuous master timecode:** Captured video uses SMPTE timecode (Control Clock set to UTC).

- **Stay within the range of common frame rates of 24 - 30 frames per second (fps):** Video is captured at 29.97 fps

Advice for File Archivists

- **Move video files off internal camera data storage, videotape, optical media or other unstable physical carriers to more stable storage media as soon as possible:** Video is duplicated across two shipboard SAN arrays that are each configured as RAID5 to guard against data loss. In addition, at the end of a season video is offloaded from the ship and stored shore side on an additional SAN. As of FY2013 OER became involved in a NOAA pilot program to address the issue of large volume archive of these datasets in a near line access model. As of FY14 this project is ongoing.

Advice for File Creators and File Archivists

- **Avoid multiple compressions and decompressions steps:** The source video is captured as ProRes at 145Mbps and stored as the same.

- **Select video encoding and wrapper formats that are well-supported now and future focused:** QuickTime and ProRes are widely used, particularly in the postproduction environment.

- **Select video encoding and wrapper formats that are supported by downstream applications:** QuickTime wrapper is readily ingestible by non-linear video editors (NLVE) such as Adobe Premiere and Apple Final Cut Pro in post-production editing.

- **Select video formats that are standardized and well-documented:** Both QuickTime and ProRes are established and well documented

- **Select formats that can contain and label complex audio configurations including multiple channels and sound fields beyond mono and stereo:** The format used by the *Okeanos* program supports 1 video stream and up to 4 audio channels (all utilized in the EVS clipping process).

When Following the Recommended Practices Is Not Practical

The project goals required some compromise and preclude following several Recommended Practices most notably in the choice of lossy compression:

- **Select uncompressed video encoding over compressed encoding:** When capturing video shipside, one of the biggest limitations that must be considered is the available storage capacity. Due to the finite physical space aboard ship, the program strives to maximize the storage capacity in the space available, but there is an upper limit depending on the storage density of available drives. As such tradeoffs must be made. In this case shipboard videographers attempt to gather as much subsurface video as possible in as high a bitrate as is operationally feasible. Thus for now and well into the foreseeable future compression for ROV footage will be a requirement.

- **If compression is used, select mathematically lossless compression over visually lossless or lossy compression:** ProRes is a lossy format. Again by necessity a balance between quality and available storage capacity had to be evaluated. Part of the reason that ProRes was preferred is that it only supports i-frames. Thus, although a lossy format, each individual frame of the video is independent of proceeding or subsequent frames to determine its individual pixel composition.

- **Select video encoding and wrapper formats that are non-proprietary:** QuickTime and ProRes are proprietary formats, but there are active open source projects actively supporting both.

LESSONS LEARNED

File Management and Data Storage

The greatest challenges in capturing video at up to 6,000 meters below the ocean surface are the engineering tasks inherent in the development of a platform that can physically go into that environment and route video back to the surface. The second biggest challenge is figuring out how to store, manage, curate, and archive the data collections given the large data volume, combined with shipboard storage and restrictions in data transmissions. In some ways it seems like the first challenge is almost the easiest. With limited digital storage space, a mechanism had to be developed to capture video streams simultaneously from multiple sources; save the data to a format that would be

easily ingestible and reusable by end users; and name and document the files in a method that would make the footage geospatially usable.

Finally the footage needed to be of a quality that would be sufficient to capture the detail necessary for scientific analysis as well as broadcast source material, but still at volume sufficient to support shipboard redundancy within finite storage. This is a problem that has been addressed - with various implementations and with varying degrees of success - with any ocean going ROV operation.

By adhering to a born digital implementation utilizing a high quality NLVE accessible format, and by developing and enforcing a strict naming convention and documentation routine, the *Okeanos Explorer* program has taken an aggressive approach to collecting and managing this dataset.

Archival Requirements

Collecting and managing born digital video at this quality generates a significant quantity of data. The ROV system on the NOAA Ship *Okeanos Explorer* is the first government owned dedicated deep-water system to produce annotated video, in this format, and at this volume. As such, the program is pioneering both scientific exploration as well as new data management capabilities in the born digital archive paradigm.

In 2002 the National Oceanographic Data Center (NODC) completed a scientific appraisal of OER data assets, and established responsibility for the archival of video data. NODC developed the Video Data Management System (VDMS) to preserve analog video assets (physical media) within the NOAA Central Library. The system includes a video lab where end users may view and copy analog materials. A small subset of digital video (both digital versions of analog data and born digital data) are available online for public review and are backed up to tape. In general, the VDMS has not adapted well to the industry transition to born digital data collections. Long term preservation of these critical environmental data assets has not been fully achieved.

NODC and OER are investigating the use of Cloud services for data storage and retrieval. The NOAA Comprehensive Large Array-data Stewardship System (CLASS) is a dedicated system for the long-term archival and distribution of NOAA environmental data. To date, CLASS is not used for digital video archival; in FY14 NODC and OER will pilot this capability.

VOICE OF AMERICA
SUSTAINABLE METADATA CAPTURE IN A BORN DIGITAL PRODUCTION ENVIRONMENT

CASE HISTORY SUMMARY

This case history describes how Voice of America approaches the requirement of ensuring descriptive metadata is associated with each house-shared asset in the digital production system and every asset in the digital archive.

CASE HISTORY AUTHOR

Pamela Commerford (pcommerford@voanews.com), Media Asset Management Branch Chief, Voice of America

INTRODUCTORY INFORMATION

Institutional Background

The Voice of America[16] (VOA) a dynamic multimedia broadcaster funded by the U.S. Government, broadcasts accurate, balanced, and comprehensive news and information to an international audience. VOA reaches our audiences via satellite delivered radio and television programming, the internet, mobile devices, Facebook, Twitter and other social media platforms using the medium that works best for specific audiences.

The Central Production Services Division is a support element within VOA which facilitates production activities of over 43 language services. The part of the division responsible for managing shared collections of video and audio, the Media Asset Management Branch, consists of approximately 30 staff members. Of the 30 members, roughly 10 staff members respond to requests for video and archive video; 10 staff members add metadata to live feeds; 5 members process and respond to audio requests and 5 members work with the physical print, audiotape and videotape collections.

Project Background

Efforts to centralize the archiving of digital video assets produced and used by the language services and central newsroom gained momentum in late 2012. By this time, the Media Asset Management (MAM) system had gone through been upgraded several times, upgrades, the template documentation was finalized and production and archiving workflows were firmly in place. By centralizing the process of Central archiving we were a able to meet our goals of providing a greater ability to search and share the assets, audit production output, and adapt to a new requirement – respond to U.S. domestic requests for VOA programs output under the Smith Mundt Modernization Act of 2013.[17]

In 2013, the Voice of America completed the migration away from a tape-based production workflow to a completely file-based workflow during the production process for digital content. VOA utilized s several software and hardware systems to manage our digital assets. The centerpiece of our system is the Dalet Plus News Management System which is connected to two Omneon Spectrum playout servers (one for studios and one for TV Master Control) for asset archiving. Dalet operates with the Front Porch Archiving system and Spectra Logic LTO storage. Digital files (and metadata when available) are imported into watch folders, travel through the video production process workflow, and play to air from a TV Studio control room through TV master control automation. The watch folders contain computer programming scripts or agents that transcode content into the house standard, DV25, and into specific folders (called Categories) in the Media Asset Management (MAM) system.

---

[16] http://www.voanews.com/
[17] http://www.bbg.gov/smith-mundt/

The focus of this paper is on managing asset metadata, from the time to asset is created to when it is sent to the archive. While the metadata exists in the MAM database rather than as a media wrapper, metadata is linked to an associated video file via unique system identifiers and is used to search and display the video file. Archivists push the assets to a separate digital archive selectively after confirming the AMF is filled out properly and selection criteria are met.

CASE HISTORY DETAILS

One uniform Asset Manager Form (AMF) is used for all assets. Different fields are utilized based on the asset type (e.g. stock footage, edited story, original field shoot, television script). The AMF contains roughly 80 unique fields, grouped in modules that can be displayed (open or closed) depending on the needs of the user. While there are ten mandatory fields for any asset to be sent to the archive, a typical asset may only have additional five or six fields to be considered fully cataloged. Assets may have only two or three fields of the AMF automatically populated, making its use-value limited to the present time. This type of item is purged automatically based on the folder purge date. Development of the AMF has been a years-long process, and the form continues to be modified as needed. Addition to or deletion of fields is heavily vetted with stakeholders to ensure all user requirements are considered.

Different types of assets have different associated information and different ways the data is populated in the fields.

1. A **Program Master**, for instance, will have summary data about the finished TV program all contained in the script field, because it is free form and can hold a practically unlimited amount of data.

2. A story received from a **newsfeed** provider will contain data in source fields, rights restrictions, keywords, issue date, suggested script, technical information, etc.

3. A **live event** gets detailed descriptions of sound bites and visuals when logged and tagged for immediate granular searching.

4. **An originally shot interview** will use entirely different fields to display specific information about the guest being interviewed, language, location, and event date.

5. An **edited video story** will have detailed script and other pertinent information peppered in a variety of fields.

In order to be able to share assets effectively, and maximize re-use value, the data must be created and entered or matched with the appropriate video file at the same time or soon after the asset enters the production system.

VOA Asset types and evolution of metadata treatment

1. Program Masters

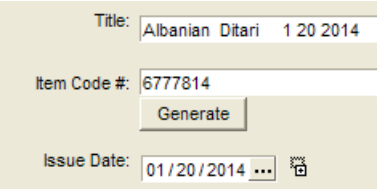
November 9, 2012 marks the date that the Video Library began efforts to centrally archive every VOA TV Program. Since the VOA Language Services was responsible for maintaining their own program masters up to this time, Video Library staff first had to assemble a listing from the Broadcast line up / current Program Catalogue using MS Excel. The data had to be sorted by Division / Language Service / title. As we reached out to Services we added Producer contacts and phone numbers. The project began with eighty-six titles that air on a daily, weekly or monthly basis.

We held several meetings with our IT department and stakeholders to establish requirements and basic workflow. All studio programs would have to be recorded, and manually placed in an archive folder by our Master Control department, in the digital system. Pre-recorded programs are held within “place holders” areas in the system, and then aired based on the broadcast schedule. In both instances, a Master Control Technician is needed to manually identify the show after it has aired, and move the file to a “to be archived” folder in the production system.

Initially, a Program Master has no descriptive data other than basic title and system generated data. Sending Program Masters to the Archive with such limited data is not an option, since the Agency needs to be able to search on titles, summaries and guest information for a variety of reasons. So, with the support of management, we enlisted program producers to add a couple of keystrokes to an already existing workflow that exists outside of our MAM system.

Program producers routinely fill out a broadcast log after each program, describing guests, topics, stories and any other show details. The additional keystrokes we implemented were: 'copy' the broadcast log data, and 'paste' it into the AMF of the Program Master. We use the script field in the AMF since this is a long-text field. Ideally this is accomplished within the first few days of airing. An archivist then reviews the metadata for each title, spots checks the actual video, and sends it to the Archive. Archivists remind the responsible Language Service producers to add the data after five days if it hasn't yet been added. Without this vital information, the Program gets purged from the system and is unavailable for future use or reference.

Asset Manager Form (AMF) fields essential for Program Masters



Title: title is automatically populated by the MAM system. Producers add 'air date' to the title for easy reference for search and sort in the Archive.

Item Code: a unique system generated identifier that is used to find the asset or further ID the asset.

Issue Date: a searchable sortable field that represents 'air date' for program masters.

Asset Manager Form (AMF) fields essential for Program Masters continued: Photo 2

SCRIPTS Field: The broadcast log is pasted into the 'scripts field' of the AMF. Using the Scripts field for this information is user friendly and as efficient as possible for this manual process.

Scripts: DITARI

Service :	Albanian	Wash. Signon :	12:00:00 PM ET	Editor :	KONDA							
Date :	01/20/14	Wash. Signoff:	12:29:15 PM ET	Producer:	LULUSHI							
Signon :	17:00:00 UTC	Target Signon:		Engineer:	LARRY							
					MC:	KONDA						
#	Number	Adaptor	Title	Voice	A	Min:SC	L/T					
1					OPENING & HEADLINES		MC	Y	00:35	T		
2					BILLBOARD	MC	Y	00:45	L			
3	1		LK		NEWSCAST 6773269 SYRIA 6773185 IRAN NUCLEAR PKG APTN - EU MINISERIAL MEETING UKRAINE PROTEST MLK DAY							
4					BUMPER: TRANSITION ID				00:10	T		
5	9		AGOLLI		STRINGER REPORT - ALBANIA ENERGY PAYMENTS				IA	Y	03:20	T
6	9		MERO-BRIEF		STRINGER BRIEF - MACEDONIA RADIKA RIVER DEVIATION/ALBANIA REACTS				MC	Y	01:00	
7	9		KOLA-BRIEF		STRINGER BRIEF - VLORA OFFICIAL WOUNDED/ CAR BLAST				MC	Y	00:50	L
8	9		MERO-BRIEF		STRINGER BRIEF - CROAT PM IN ALBANIA				MC	Y	00:45	L
9	9		ABAZI		STRINGER REPORT - KOSOVA - MITROVICA ELECTION				BA	Y	03:00	T
10	9		SHEHU		STRINGER REPORT - KOSOVA PARLAMENTARY REFORMS				LS	Y	03:30	T
11	6771465		ZE		Washington Week Ahead		ZE	Y	02:40	T		
12	6753087		JA		ML KING / BIRMINGHAM SAFE HOUSE			JA	Y	03:10	T	
13	6698642		JA		ISRAEL ART HOLOCAUST			JZ	Y	03:40	T	
14					CLOSING THE SHOW			MC	Y	00:10	L	00:10 L

2. Newsfeeds

The agency subscribes to commercial news services that provide video packages and unnarrated video stories covering all events of the day, international and domestic, file material and breaking news. All of the file based stories have associated metadata, but until late in 2013, the television script files were not linked or easily associated with a video file. Manually finding a script and then matching it to the video in the MAM was time consuming and frustrating, often compounding producer deadlines. Matching video must be found before it can be viewed, and titles are often extremely similar depending on a news story. In addition to immediate use, stories are selectively archived for potential future use of the video. In order to be effectively captured; metadata must be fully linked to the video from its first appearance in the MAM.

Over the course of the year we grew in our ability to automate the process of script marrying. XML-provided script information is funneled through an XSLT translation script in the television script Watch Folder, to populate to mapped fields in the AMF with the appropriate data.

In the following illustrations, you can see a small portion of the original XML, select lines of XSL template, and the resulting XML. The XSL is a standardized "translation tool," that can be customized to the specific requirements of the task at hand. This tool was established many years ago and continues to be refined by the World Wide Web Consortium (W3C), an "international community that develops open standards to ensure the long-term growth of the web". Use of the translation tool in our MAM environment relieved staff of manually matching hundreds of scripts to video files daily. Each XSL must be custom designed for each newsfeed service we subscribe to, and tweaked when the newsfeed provider makes any adjustments to the distributed newsfeed product.

A small clip of a much longer original XML – data you can see includes the headline:

```
    </genre>
    <by xml:lang="en">Reuters, FEB 07</by>
    <slugline xml:lang="en">SPAIN-ROYAL FRAUD (O)</slugline>
    <headline xml:lang="en">Spain's royal fraud scandal</headline>
    <dateline>FEBRUARY 07, 2014</dateline>
    <description role="descRole:videoCaption" xml:lang="en">5014BO-SPAIN-ROYAL_FRAUD_O_</description>
</contentMeta>
```

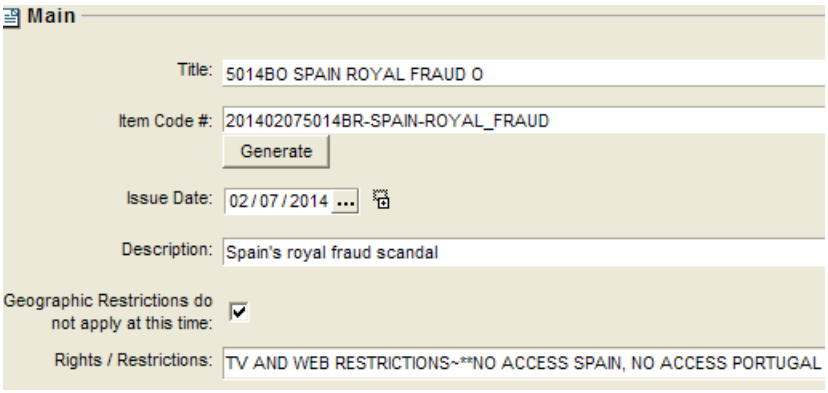
Select lines of a much longer XSL – this tool labels what goes where, such as the ***headline*** *being targeted for the* ***Description*** *field and* ***usageTerms*** *for* ***RightsComments3****.*

```
...
<xsl:template match="iptc:packageItem">
    <key1><xsl:value-of select="substring-before (iptc:itemMeta/iptc:fileName,'.')"/></key1>
...
    <Description><xsl:value-of select="iptc:contentMeta/iptc:headline"/></Description>
    <RightsComments3><xsl:value-of select="iptc:rightsInfo/iptc:usageTerms[position()=1]"/></RightsComments3>
    <Title_subname><xsl:value-of select="iptc:itemMeta/iptc:fileName"/></Title_subname>
    <Rights><xsl:value-of select="iptc:rightsInfo/iptc:usageTerms[position()=1]"/></Rights>
...
</xsl: template>
```

A small clip of the resulting transformed XML – here you see some of the fields that are populated as a result of the XSL:

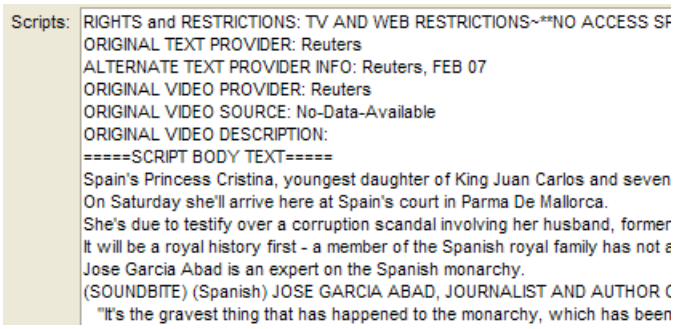
```
<?xml version="1.0" encoding="UTF-8"?>
<Titles xmlns:iptc="http://iptc.org/std/nar/2006-10-01/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance" xmlns:rtr="
http://www.reuters.com/ns/2003/08/content" xmlns:html="http://www.w3.org/1999/xhtml">
<Title>
<Rights2>TV AND WEB RESTRICTIONS~**NO ACCESS SPAIN, NO ACCESS PORTUGAL</Rights2>
<key1>201402075014BR-SPAIN-ROYAL_FRAUD</key1>
<title>5014BO SPAIN ROYAL FRAUD O </title>
<IssueDate>2014-02-07</IssueDate>
<Description>Spain's royal fraud scandal</Description>
<RightsComments3>TV AND WEB RESTRICTIONS~**NO ACCESS SPAIN, NO ACCESS PORTUGAL</RightsComments3>
<Title_subname>201402075014BR-SPAIN-ROYAL_FRAUD.xml</Title_subname>
<Rights>TV AND WEB RESTRICTIONS~**NO ACCESS SPAIN, NO ACCESS PORTUGAL</Rights>
```

Resulting AMF:



All of the fields here are populated from an XML script provided by a newsfeed service.

Our IT staff worked closely with the MAM vendor to build the XSLT translation program within a Watch Folder, and to populate automatically the data into the AMF metadata fields of the matching video asset.



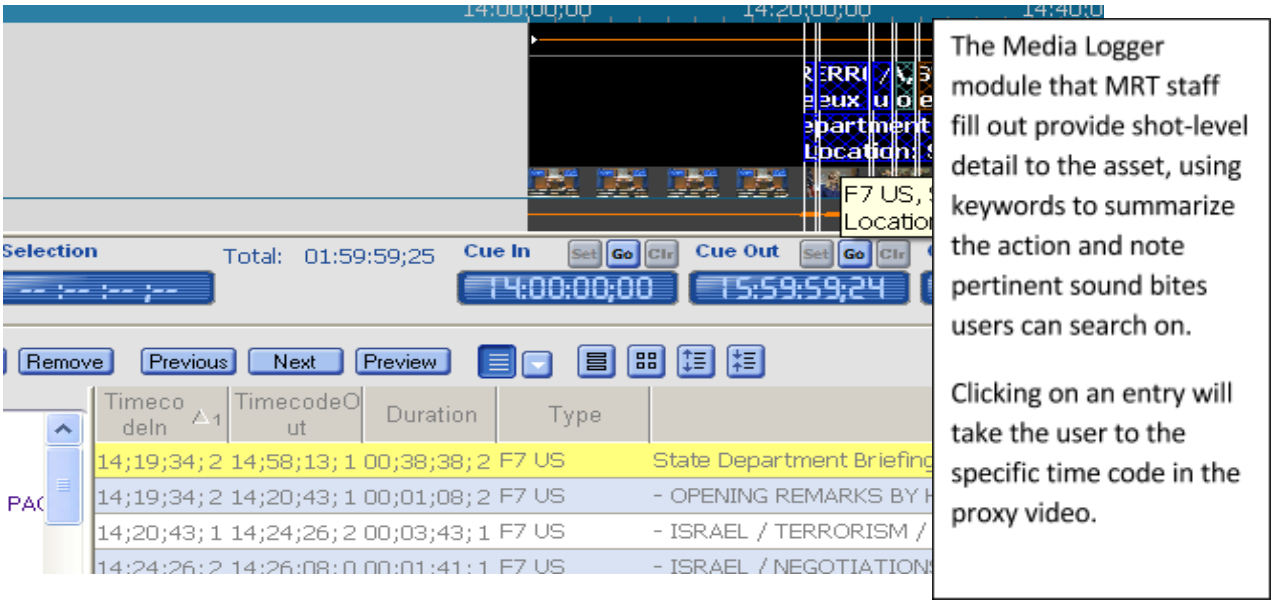
The entire script populates the scripts field, so users can easily copy this info and paste it into an e-mail or application outside of the MAM.

Fields also populate by system default settings in the watch folder the assets pass through before entering the MAM.

3. Live Feeds

VOA records live event feeds on a 24 hour – 7 day a week basis. Starting in 2009, we began centralizing the process of logging and tagging of live event feeds for shared VOA use. The Media Resource Team (MRT) now operates with up to ten members and is staffed to log and tag live feeds 12 hours per day on weekdays, and 8 hours on weekends and holidays. The MRT is scheduled around the prime times that producers use the feeds, and make live events keyword searchable within seconds of the addition of the metadata. MRT staff use a different component of the MAM system – the Media Logger, to log shot-details. The AMF is also filled out, and when a transcript is available it is added to the 'transcript' field of the AMF. MRT coverage is extended for special events coverage, such as the President's State of the Union Speech, elections and other high interest events or breaking news. With such rich metadata, much of the work of the MRT makes it into the archive.

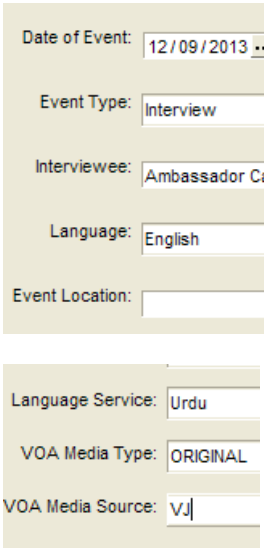
Media Logger of Asset (another component of the system that offers keyword searching)



4. <u>Original Ingest</u>

In order to capture and make VOA original content available for shared and long term use, producers are encouraged to bring high value field video to central ingest. Two different elements coordinate response to ingest requests in order to ensure a wide range of support-times for quick turnaround. Immediately after ingest, the video is logged and tagged by the MRT and an announcement sent VOA-wide. Since not all original video needs to be shared nor archived, ingesting into the MAM remains optional. Producers upload video to storage outside of the MAM, which typically remains at capacity. Since this is an unmanaged storage system, with no attached metadata other than filename, leadership support is essential for moving toward the centralized, long-term solution of the MAM.

Example of AMF of VOA Original asset



AMF fields specifically for use for ingested original material. The producer fills out a form with pertinent data. The ingestor enters the title and other basic data.

The asset then goes to the MRT to log and tag the video using the Media Logger component of the MAM.

5. Central News stories and other VOA produced video

Our production workflows are built around the core idea that stories can be shared across the language services. Our Central News room produces current news stories that can be re-tracked by any language service for broadcast within hours. Stories include originally shot video as well as newsfeed video. Producers write scripts and add source and other information to the AMF attached to the script. The video is produced separately, and is embedded into the script and published. Since a separate AMF exists for the video, it often doesn't contain the rich data of the textual script. Producers select stories to send to the archive, and copy and paste info from the script AMF to the Video AMF. The video is then sent to the archive, with its data-rich AMF.

SUMMARY OF VIDEO AND AUDIO DATA TECHNICAL SPECIFICATIONS

VOA employs a "house style" for all its broadcast deliverables, including those outlined in this case history project.

Video Data

- Container/wrapper: QuickTime (.mov)
- Target total bitrate: 25 Mbps
- Frame size: 720 x 480
- Aspect ratio: 4:3
- Video]codec: DV25
- Compression type: Lossy
- Video frame fate: 29.97 fps
- Color encoding: YUV
- Chroma format: 4:1:1
- Interlaced scan

Audio Data

- Audio Channels: 2
- Audio Codec: PCM
- Audio Sample Rate: 48 kHz
- Audio Bit depth: 16 bit

RELEVANT RECOMMENDED PRACTICES

The project goals permitted following many of the Recommended Practices including:

Advice for File Creators

- **Provide the means to collect and submit metadata starting at the video shoot:** VOA-MMAM uses its custom built AMF (Asset Manager Form) with fields tailored for each asset type
- **Stay within the range of common frame rates of 24 - 30 frames per second (fps**): VOA-MMAM uses the standard 29.97 fps

Advice for File Archivists

- **Retain all the data from the original file if the video file structure is changed:** VOA-MMAM case history touches upon associated and embedded data

- **Move video files off internal camera data storage, videotape, optical media or other unstable physical carriers to more stable storage media as soon as possible:** VOA-MMAM uses the Dalet Plus News Management System which is connected to two Omneon Spectrum playout servers (one for studios and one for TV Master Control) for asset archiving. Dalet operates with the Front Porch Archiving system and Spectra Logic LTO storage.

Advice for File Creators and File Archivists

- **Use XML-based metadata schemas with strong support for digital video:** VOA-MMAM uses standard XML-compliant metadata templates which evolved over time and were adjusted as needed throughout the project.
- **Document and use technical metadata:** VOA-MMAM utilizes a sophisticated workflow to using an XSLT translation script to populate to mapped fields in their customer Asset Manager Form with the appropriate data.
- **Select video encoding and wrapper formats that are supported by downstream applications:** VOA-MMAM has a controlled internal system so all formats are well supported.

When Following the Recommended Practices Is Not Practical

The project goals required some compromise and preclude following several Recommended Practices:

- **Select High Definition (HD) video encoding over Standard Definition (SD):** VOA-MMAM, following the house style, uses SD.
- **Select higher bit rates over lower bit rates:** VOA-MMAM's internal house standard is 25 Mbps. which meets VOA-MMAM's business needs and is supported by essential internal VOA-MMAM systems.
- **Use higher chroma subsampling ratios rather than lower:** VOA-MMAM's internal house standard is 4:1:1 which meets VOA's business needs and is supported by essential internal systems.
- **If compression is used, select mathematically lossless compression over visually lossless or lossy compression**: VOA-MMAM's house standard is DV25. While DV25 is a lossy codec, it meets VOA-MMAM's business needs and is supported by essential internal VOA-MMAM systems.

LESSONS LEARNED

The First Step is Always the Hardest

Transitioning from a videotape-based world of TV production to all-digital TV production has been a paradigm shift for the Voice of America. For quite some time, basic navigation skills and workflow development competed with our regular duties and deadlines. As the MAM project team worked tirelessly to address issue after issue and VOA staff learned to use the tool better, resistance turned to change and change turned to adoption. With enhancements and customization of the software, we are more able to manage the urgent production deadlines and initiate steps toward future goals. The video library's attention at this point is turned to fine tuning the details – looking for more tools to enhance metadata matching and more efficient ways to manage the flood of content. Production has already shifted to a new house standard video aspect ratio, 16x9; and the transition to HD is on the horizon.

The Tools are Out There

Descriptive metadata can be captured early in the life cycle of video, and metadata standards are more widespread and shared than ever. Distributors and suppliers can and do use metadata standards that can be parsed and put back together to automatically feed a MAM. The first question to ask is 'What is the standard being used?' the second question: 'how can it be transformed, fed and matched automatically to video in our MAM system.'

It Takes a Human Eye

A week did not go by without the team discovering the unexpected, or learning something new, or having to adjust a workflow or expectation. There are so many possible points of failure when working with complex systems, multiple teams and the expanding universe of options. Whether it was drifting timecode, a vendor adjusting a seeming minor delivery procedure, a new contract to set up, a camera with a switch set differently than others, it was often discovered in the moment. The 24-hour System Monitoring / response team saved the day many-a-time.

Once A Metadata Schema Is Adopted, Time and Use Will Determine the Value of the Fields Chosen

We found that our main metadata tool – the Asset Manager Form – continued to evolve even after a great deal of time and energy was spent by all the stake holders to create it. Actual use put it to the test. Flexibility in this area with good communication to users on changes occurred throughout our efforts.

Automate – Automate - Automate

In our implementation, system defaults in the MAM system have been set to auto-populate fields based on the folder in which they exist. This removes the need for repetitive human data entry, and ensures designated assets move along the chain to the proper categories or reflect consistent data applicable to that series of assets. We continue to look for other parameters and ways to auto-populate descriptive fields including XML and XSL availabilities.

It Takes a Village

From beginning to end of the life cycle of an asset, many hands must not only touch it but leave their mark. The question "what should I archive," must be framed in light of the ability to associate descriptive metadata with the essence, rather than with an eye toward archiving "everything possible." If there is not enough commitment from the start, then creating a successful framework for archiving valuable content is difficult to realize.

RECOMMENDED RESOURCES

XSL Transformations (XSLT) Version 1.0 W3C Recommendation 16 November 1999. http://www.w3.org/TR/xslt

International Press Telecommunications Council. http://www.iptc.org/site/Home/

http://www.iptc.org/std/catalog/catalog.IPTC-G2-Standards_3.xml This file is provided as is by the International Press Telecommunication Council, IPTC - www.iptc.org

ARCHIVING BORN DIGITAL VIDEO CASE HISTORIES

The five *Archiving* case histories tell the story of bringing the born digital video files into managed data repositories for long term retention and access. These case histories explore the issues which emerge when the born digital video objects arrive at the archive. They cover topics including the challenges of dealing with diverse formats, understanding and documenting relationships among the video files and related objects, and metadata. A major topic for this case history sets is the technical characteristics of file formats: how to identify and document what formats comes in to the archive, when are changes to the file attributes needed, and what are the impact of changes to the format and encoding.

Each case history focuses on a different problem set. LC-NAVCC-VEF describes the rationale for normalizing all video files to a single evergreen format prior to repository ingest. LC-WebArch-YouTube describes the challenges of harvesting digital video from YouTube. NARA-BRCC addresses the complexity of making bundled heterogeneous files ready for repository ingest and access. SIA-DVD project details the process of harvesting video files off authored DVDs. SI-DAMS describes the complex challenges of bringing digital video files and related objects into a centralized digital asset management system.

LIBRARY OF CONGRESS, PACKARD CAMPUS OF THE NATIONAL AUDIO-VISUAL CONSERVATION CENTER: THE CASE FOR NORMALIZATION TO AN EVERGREEN FORMAT

CASE HISTORY SUMMARY

The Library of Congress' Packard Campus of the National Audio-Visual Conservation Center is the state-of-the-art facility which acquires, preserves and provides access to the world's largest and most comprehensive collection of films, television programs, radio broadcasts, sound recordings, and software and electronic gaming and learning. The scale of the collection is massive so the need to define a single target moving image normalization format, JPEG2000 lossless in MXF OP1a, is well justified in order to maintain interoperability within workflow processes and long term sustainability of the archived files.

CASE HISTORY AUTHOR

James Snyder (jsny@loc.gov), Senior Systems Administrator, NAVCC, Library of Congress

INTRODUCTORY INFORMATION

Institution Background

The Packard Campus of the National Audio-Visual Conservation Center at the Library of Congress[18] is the a state-of-the-art facility which acquires, preserves and provides access to the world's largest and most comprehensive collection of films, television programs, radio broadcasts, and sound recordings. Located in Culpeper, Virginia, through a partnership with Packard Humanities Institute, the United States Congress, the Library of Congress, and the Architect of the Capitol, the Packard Campus hosts 415,000 square feet, more than 90 miles of shelving for collections storage, 35 climate controlled vaults for sound recording, safety film, and videotape, and 124 individual vaults for more flammable nitrate film. The Campus has globally unprecedented capabilities and capacities for the preservation reformatting of all audiovisual media formats (including obsolete formats dating back 100 years) and their long-term safekeeping in a multi-petabyte-level digital storage archive. In addition to preserving the collections of the Library, the Packard Campus was also designed to provide similar preservation services for other archives and libraries in both the public and private sector.

Collection Background

Collections at the Packard Campus consist of two broad categories. The first group, comprising approximately 80% of the physical collection and a significant portion of the born digital collection, is copyright deposit for specific content areas: audio, video, motion picture film, and electronic gaming and learning. The second group of material with the Packard Campus are the non-copyright deposit collections. Significant collections in this group include the 1950s gift of nitrate films from five major movie studios, the Afghan Media Resource Center collection which includes original interviews and raw footage from over 30 years of war in Afghanistan, the HistoryMakers African American video oral history project, the American Archive of Public Broadcasting historic collection of American public radio and television content, and many others.

The entirety of the Library's audiovisual holdings will be digitized, creating both archive masters and access copies providing researchers with playback on demand in the Library's Capitol Hill reading rooms. The digital archive is based on the concept of data migration and verification. Migration to progressively higher density storage—meaning progressively greater storage capacity at the same or lower cost for as long as these economies exist—will continue indefinitely into the future. The NAVCC media data storage system is automatically checked on a periodic schedule using SHA-1 cryptographic hash checksums. This decision was made because circumstantial experience has shown

[18] http://www.loc.gov/avconservation/packard/

that physical movement of tapes outside of a data robot environment meaningfully increases the data corruption rate of the contents stored on the tape.

All digital material, whether submitted as a digital file or captured to file at the Library, is normalized to a standard file format configuration. For moving image content, this is JPEG2000 lossless in MXF OP1a.[19]

The retention period for copyright deposit is 150 years so the digital material in the archive is essentially a permanent data set. The archive currently contains about 5 PB and 200K archive files of digital and digitized audio and moving image recordings.

CASE HISTORY DETAILS

The case history explores the rationale of normalizing a diverse set of heterogeneous born digital video file formats and encodings into just one evergreen format for long term preservation which encourages longevity despite inevitable technology changes.

While this case history focuses on digital video within the collections housed at the Packard Campus, the same principles apply to film and audio collections. For reference, the evergreen formats for these collections are:

Audio Evergreen Format Details

- Broadcast WAV (BWF RF64) 96 kHz/24 bit
- Digital audio captured at native sampling & bitrates IF they are AES standard
- Access proxies: WAV 44.1 kHz/16 bit
- Converted to MP3 or other format if necessary

Digitized Film Evergreen Format Details

- DPX with BWF RF64 audio: current limited production needs & preservation testing
- Aiming for JPEG2000 Lossless in MXF OP1a
- Access proxies HD MPEG-4 H.264

Format Sustainability through International Standards

The Packard Campus receives digital video files through a variety of input streams including copyright submission and general collection acquisition. This diversity in sources also leads to wide heterogeneity in and among file types and attributes. In order to reduce the variability in these large and complex collections, all digital video inputs are normalized on ingest to one standard format, JPEG2000 in MXF OP1a. The benefit of normalizing on ingest into the repository is that the toolkits are still (mostly) available for the submitted file formats and encodings. After a time, general production and toolkit availability will decrease and it may not be possible to access and transform the file.

Why JPEG2000 encoding in the MXF OP1a wrapper? One reason is that both are international and well-documented standards and preservation of content in a given digital format over the long term is not feasible without an understanding of how the information is represented (encoded) as bits and bytes in digital files. JPEG2000 is standardized in ISO 15444. MXF OP1a is standardized through SMPTE 377-1 and SMPTE 378M-2004. A compelling benefit of adopting internationally standardized wrappers and codecs is that vendors build tools and application to meet these specifications because that will help wide market adoption. Another reason is that preservation planning is simplified because institutions only need to maintain the documentation on one set of standards instead of many.

The version of JPEG2000 adopted by the Packard Campus[20] is the mathematically lossless compression level reversible 5/3. (Lossy JPEG2000 profiles are also available but the Packard Campus implements a lossless compression profile.) This means that the compression is completely reversible and there is no loss of quality when the file is encoded and decoded. Other attractive features include that it does not have licensing issues, it can be

[19] http://www.digitalpreservation.gov/formats/fdd/fdd000206.shtml
[20] http://www.digitalpreservation.gov/formats/fdd/fdd000206.shtml

wrapped in a standardized file wrapper (MXF) which promotes interoperability, it can accommodate any color spaces (YPbPr, RGB, XYZ, and new ones being worked on) and it can accommodate any bit depths (Video: 8 & 10 bits/channel; Film: 10, 12, 16 bits/channel). The Packard Campus plans to maintain the bit depth for source images beyond 10-bits.

The Packard Campus currently uses 10-bit for video because this matches the specifications laid out for serial digital interface as defined by SMPTE starting with ST 259M (although as stated above, NAVCC is preparing to accommodate native bit depths beyond 10-bit). Additionally, 10-bit encoding is preferred over 8-bit as a harmonization encoding so that decoder software writers do not have to accommodate both.

The MXF wrapper was specifically designed to aid interoperability and interchange between different vendor systems, especially within the media and entertainment production communities which are the primary content providers to Packard Campus collections. The file specification was standardized by the SMPTE (Society of Motion Picture & Television Engineers) and AMWA (Advanced Media Workflow Association) and allows different variations of files to be created for specific production environments and can act as a wrapper for metadata & other types of associated data.

Other possible file specifications are not viable for a number of reasons. AVI, for example is limited by being vendor specific and an aging toolset. MOV is in widespread use but it is proprietary, requires license fees and is not well documented. MPEG-2 and MPEG-4 are standardized toolkits but not well-documented coding standards that are repeatable time after time. The ISO/IEC 13818 documentation set, for example, defines what the structure of the file should be but doesn't specify how the essences are encoded. The MPEG family in general allows for a large number of variations in how each file can be encoded and decoded and it's impossible to document all the possible variations since the toolkit implementations are different between vendors, and many times between versions of the same vendor's software code.

Rationale to Use Lossless Compression

The Packard Campus implements a reversible and mathematically losslessly compressed profile of JPEG2000 in order to save digital storage space and by extension, cost. The JPEG 2000 core coding, lossless compression scheme documented in ISO/IEC 15444-1:2004 averages 2.5 to 1 compression ratio. With collections at the scale of those at the Packard Campus, this translates into considerable savings in digital storage space and costs compared to storing uncompressed files. Digital storage costs tend to decrease as technology advances and capacity increases but the volume in the Library of Congress collections still justifies keeping the needed space. In addition, lossless compression does not significantly add to processing complexity over uncompressed files because the more data present, the more complex the processing. The savings in the volume of data justifies the complexity of the file structure.

Interoperability

Interoperability is the extent to which systems and devices can interchange data. It is essential in high volume file-based media preservation workflows in place at the Packard Campus and for the exchange of files between institutions and collections. In short, systems and devices from different vendors need to be able to successfully pass data among each other. By limiting the format variables to just JPEG2000 in MXF OP1a, the Packard Campus is able to select features that are not vendor specific. Adopting this internationally standard file configuration means that there are more options available in the commercial vendor market. The Packard Campus is able to use as many off-the-shelf products as feasible and invent items or write custom software only when a commercial product does not fit the need.

Abundant, Robust and Managed Archival Storage is Essential

The digital archive at the Packard Campus is able to support the very large quantities of digital objects (not just video but all types of content) within the collections. As of this writing, the archive contains about 5 PB of recorded data with 25 PB are available and about 200,000 moving image and audio archive files. It is currently operating at about 20% of designed production levels so there is ample room to grow.

- Dual copies, geographically dispersed
- Oracle Sun StorageTek T10000-C data tapes

- 9800 slots, 4900 currently populated.
- SHA-1 Cryptographic checksum used to verify integrity of files from the point of creation to the archive as well as to verify the integrity of the archive over time at regular intervals
- Metadata is maintained in databases
- Proxies are maintained on servers, and written to the archive

SUMMARY OF VIDEO AND AUDIO DATA TECHNICAL SPECIFICATIONS

Video Data

- Container/wrapper: MXF OP1a
- Target total bitrate: average of 90 Mbps for SD; 800 Mbps HD (uncompressed HD input)
- Timecode: native
- Frame size: native
- Aspect ratio: native 4:3 for SD video or 16:9 for HD video
- Video codec: JPEG2000 Lossless reversible 5/3
- Compression type: Mathematically lossless
- Video frame rate: native
- Color space: native
- Chroma sampling format: native 4:2:2 for YPbPr, 4:4:4 for RGB
- Bit depth: 10-bit (and native beyond 10-bit)
- Interlaced/Progressive scan: native

Audio Data

- Audio channels: native
- Audio codec: PCM
- Audio sample rate: 48
- Audio bit depth: 24
- Note: If timecode is stored in an audio channel, it is captured as an audio channel and duplicated in a timecode track

RELEVANT RECOMMENDED PRACTICES

Advice for File Archivists

The use of JPEG2000 lossless encoding reversible 5/3 in MXF OP1a as the normalized target format supports a wide range of Recommended Practices including:

- **If compression is used, select mathematically lossless compression over visually lossless or lossy compression:** The profile of JPEG2000 used in the Packard Campus Evergreen format is reversible and mathematically losslessly compressed.

- **Select video encoding and wrapper formats that are well-supported now and future focused:** JPEG 2000 lossless and MXF are both well supported in current toolsets and are likely to continue to be so.

- **Select video encoding and wrapper formats that are non-proprietary:** Both JPEG 2000 and MXF are international standards.

- **Select video encoding and wrapper formats that are supported by downstream applications:** Because NAVCC normalizes all video data to the same encoding and wrapping formats, all toolsets are geared to work with these format selections.

- **Select video formats that are standardized and well-documented:** JPEG2000 is standardized in ISO 15444. MXF OP1a is standardized through SMPTE 377-1 and SMPTE 378M-2004.

- **Select video formats with capacity for robust and detailed technical metadata:** The MXF wrapper was specifically designed to aid interoperability and interchange between different vendor systems, especially within the media and entertainment production communities which are the primary content providers to Packard Campus collections. The file specification was standardized by the SMPTE (Society of Motion Picture & Television Engineers) & AMWA (Advanced Media Workflow Association) and allows different variations of files to be created for specific production environments and can act as a wrapper for metadata & other types of associated data.

- **Select video formats with greater capacity for embedded metadata over less metadata capacity:** MXF can contain detailed embedded metadata. The FADGI MXF AS-07 Application Specification for Preservation and Archiving (upcoming), which the Packard Campus has been instrumental in helping to design, can support embedded text-based data including supplementary metadata.

- **Select formats that can contain and label complex audio configurations including multiple channels and sound fields beyond mono and stereo:** LC-NAVCC-VEF uses MXF which has strong support for multiple and complex audio configurations

- **Select formats that can support robust timecode data:** MXF can support multiple timecode tracks includes a continuous Master Timecode and Historical Source Timecodes inherited from the original.

NAVCC's collection processing and archiving workflows supports the following recommended practices:

- **Identify the file characteristics at the most granular level possible, including the wrapper and video stream encoding:** LC-NAVCC-VEF uses format identification tools.

- **Determine criteria for when (if ever) it is appropriate to change the video file's technical properties (including normalization):** LC-NAVCC-VEF case history details the rationale for normalizing file types

- **Retain the original video file as submitted if transcoding, normalizing or otherwise changing the video stream to meet business needs:** Originals are retained as submitted.

- **Select appropriate technical characteristics for the video encoding if transcoding, normalizing or otherwise changing the video stream to meet business needs:** LC-NAVCC-VEF case history details the rationale for selecting file characteristics for normalizing ingested files

- **Generate a new high integrity and continuous master timecode, especially if there is no timecode in the original material:** LC-NAVCC-VEF plans to create a continuous master timecode as part of the planned MXF AS-07 implementation.

- **Retain original timecode(s) if provided, even if you generate a new high integrity continuous master timecode:** LC-NAVCC-VEF case history project retains timecode if present when normalizing files.

- **Retain the original chroma subsampling if the video data is transcoded:** Native chroma subsampling is retain when it's declared or knowable through metadata extraction.

- **Retain original frame rates if the video data is transcoded, even when they are beyond the standard 24 - 30 fps:** Native frame rate is always retained.

- **Move video files off internal camera data storage, videotape, optical media or other unstable physical carriers to more stable storage media as soon as possible:** Video data is moved off optical discs and external hard drives into managed storage as soon as it can after accession of the collection.

- **Document and use technical metadata:** LC-NAVCC-VEF reports that metadata is perhaps the area of in need of greatest focus: metadata pathways through the entire file-based workflow systems must be enabled and made reliable; metadata embedding standards need to be developed. (The FADGI AS-07 project is a big step in the right direction.); metadata schema standards need to be developed in coordination with industry

- **Create access, viewing or proxy copies with appropriate technical characteristics to meet expected use cases:** Case history details the technical characteristic of access copies.

When Following the Recommended Practices Is Not Practical

- **Select uncompressed video encoding over compressed encoding:** With collections at the scale of those at the Packard Campus, implementing compression translates into considerable savings in digital storage space and costs compared to storing uncompressed files.
- **Stay within the same codec family if the video data is transcoded:** The Packard Campus uses the house standard JPEG2000 in MXF OP1a target normalization format for all digital video files, regardless of the source format.

LESSONS LEARNED

It's Not So Hard Once You Get Started

Mass migration is not only possible, but relatively easy once you have the processes figured out.

Physical workflows (people, cataloging, movement of assets) MUST be created and implemented at the same time as the technical workflows.

We Don't Have It All Figured Out Yet

Many processes and workflows are functioning at high capacity but there is still work to be done. In early planning, it was thought that it might take about four hours to process (create metadata, deal with copyright, normalized file formats) each hour of content so that it can be made usable. In reality, it takes about 20-50 hours to process each hour of content, jumping from an estimated 4:1 ratio to up to 50:1.

Metadata is perhaps the area in need of greatest focus going forward. In short, we need more of it, we need to collect it more reliably and we need to embed it within the files themselves.

- Metadata pathways through the entire file-based workflow systems must be enabled and made reliable.
- Metadata embedding standards need to be developed. (The FADGI AS-07 project is a big step in the right direction.[21])
- Metadata schema standards need to be developed in coordination with industry.

In some areas, we know what needs to be accomplished but we need to wait for vendors to build tools which will allow us to do the work.

[21] For more information, see http://www.digitizationguidelines.gov/guidelines/MXF_app_spec.html

LIBRARY OF CONGRESS WEB ARCHIVING TEAM
PROCESSING AND REPLAYING VIDEO COLLECTED FROM THE WEB

CASE HISTORY SUMMARY

This case history serves to explore the various challenges that the Library of Congress' Web Archiving team faces in harvesting and processing large collections of YouTube video content online. We have discovered that YouTube frequently changes its backend method of serving videos, which demands adaptability in the Wayback Machine program responsible for replaying our archived content. Additionally, as YouTube content becomes more prominent in our collections, our ingest volume has expanded greatly, due to the voluminous nature of video content files. We have employed a deduplication method to combat unnecessary increases in crawl volume, but the crawls still produce sizeable outputs. This reality in conjunction with high traffic, multi-department usage of the Library's internal CTS ingest system has presented a bottleneck in our processing capacity. While these issues do present significant obstacles, we are working to remain adaptable and efficient in our procedures to balance out the limitations that we experience in replaying our YouTube content and ingesting/processing our expanding crawl outputs.

CASE HISTORY AUTHORS

Abbie Grotke (abgr@loc.gov), Web Archiving Team Lead

Gina Jones (gjon@loc.gov), Web Archiving IT Specialist

Phil Ardery (pard@loc.gov), Web Archiving IT Specialist

INTRODUCTORY INFORMATION

Institutional Background

The Library of Congress' Web Archiving team consists of 5 full time employees. Currently, the web archive spans 17 publicly available collections in addition to a continuously expanding number of working collections that are being prepped for future release. [22] Each collection has its own focus, and together they cover a wide range of cultural content both domestic and international in nature. The Library began crawling in 2000, and has since amassed over 500 Terabytes of archived web content.

Collection Background

Each collection consists of numerous crawled URIs, which are internally selected by a group of content nominators within the Library of Congress. When adding a URI to be crawled for a collection, these nominators work with the Web Archiving team to define a specific scope for the site in question. This scope ultimately functions as a set of instructions that tells the web crawler which parts—content, links, and files— of a website to scrape, and how many layers into the website the crawler should go. Currently we use the Heritrix Web Crawler. While the Web Archiving team does conduct in-house crawling, the majority of its crawling is outsourced to contractors.

Websites such as US government sites are free to be crawled without notification or permission due to their being in the public domain, but other websites might require specific permission to be crawled. Depending on the website in question, this permission may need to be explicitly granted, or may be assumed after a notice is sent to the site owner, informing him/her that the Library intends to crawl the site for a specific collection.

Given the nature of the internet, the content collected spans a wide range of formats. After each crawl completes, the resulting content is compressed into various WARC[23] formatted files. When we are performing in-house crawls, this WARC compression is performed by our in-house instance of the Heritrix crawler, otherwise it is performed by the

[22] http://lcweb2.loc.gov/diglib/lcwa/html/lcwa-home.html
[23] http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml

Heritrix crawler that our contractor—currently the Internet Archive—is running. When a user accesses the Library's web archive, a local instance of the Wayback Machine replays the necessary WARC file(s) to create the dynamic web archive experience.

Heritrix -- https://webarchive.jira.com/wiki/display/Heritrix/Heritrix;jsessionid=7AF9310B2FBECB22D3CAF140E4CCE867
Wayback Machine -- *https://webarchive.jira.com/wiki/display/wayback/Wayback+Installation+and+Configuration+Guide*
WARC File Format -- *http://www.digitalpreservation.gov/formats/fdd/fdd000236.shtml*
WARC Specifications Overview -- *http://archive-access.sourceforge.net/warc/*

CASE HISTORY DETAILS

For this case history, we would like to explore the issues of storage, in the context of mass video collection within a limited storage environment, and of replaying harvested content from a constantly evolving service provider, YouTube.

Storage & Processing

As video content is becoming more prevalent on the internet, it has naturally become increasingly relevant to our various collections. This reality coupled with the sizeable nature of video files has presented a challenge as we attempt to grapple with the ever-expanding size of our crawl outputs. As storage space costs money, budgets naturally limit the amount of content that we can store and make available. However, the more tangible issue that we currently face is a limitation regarding the speed with which we can process our content. After each crawl is completed by the Internet Archive, we must use in-house tools to migrate the content to our local network, copy it to long term storage, process it for public access, and then copy it to public access storage. The primary content transfer tool that we use is shared by many different departments within the Library, and thus experiences high traffic on a regular basis. Consequently, while the crawler may be physically capable of capturing hundreds of terabytes of content each month, we are limited by our ability to transfer, copy, and process far fewer terabytes in the same time period.

Replay

YouTube in particular provides an interesting challenge to our team, given the various and often complex methods with which the site serves up its video content. Videos are most commonly accessed through a generic youtube.com/watch?v= base URI. In playing the video, however, the site looks to various reference files which in turn point to alternate URIs where the video content itself is actually stored. Our scoping techniques allow us to successfully capture the videos when crawling the site, but the Wayback Machine piece must be configured to mimic this referential structure in order to replay the video content natively, within the archived copy of the site. Achieving the proper configuration has proved challenging in its own right, and is complicated further by the tendency for YouTube to regularly alter its methods of serving videos. The ever changing nature of this issue has presented an ongoing struggle of sorts, which we are working with the Internet Archive to address.

SUMMARY OF VIDEO AND AUDIO DATA TECHNICAL SPECIFICATIONS

Websites present videos in a wide range of different formats. Currently, YouTube accepts files in the following encoding formats:

- *.mov
- *.mpeg4
- *.avi
- *.wmv
- *.mpegps
- *.flv
- 3GPP
- WebM

In June of last year, our team worked with the Library's formats division to determine that mpeg formats should take precedence over *.flv files, when the former is available. We can analyze our archived video content by reviewing the mime-types returned by the host server when our crawler encounters the content. Each crawl yields a series of reports, including a mime-type report that provides URL and byte counts for each mime-type within the given crawl. These reports suggest that the most common video mime-types that we currently encounter are:

- video/mp4
- video/x-ms-asf
- video/x-ms-wvx
- video/quicktime
- video/x-ms-wmv
- video/x-flv
- video/mpeg

This list is based on all video captured, not just YouTube content. For example, the .flv files that we have captured could be either flash videos pulled from other (non-YouTube) sites, or YouTube videos that the crawler determined to be only available in the .flv format

For the time being, we do not have a means of producing sharply accurate figures for how much video content, YouTube content, or specific file type content we have in our archive. We are working with the Internet Archive to create an environment that will allow us to perform such analysis in the future. For the time being, we can only rely on the figures in our mime-type reports.

One shortcoming of these reports is that the crawler produces them before de-duplicating the crawl content. If the crawler encounters a video that it has already captured, for example, it will naturally harvest another capture of the video, which should be recognized as a duplicate and discarded in the post-processing phase. Thus, while the reports suggest that we have crawled around 292 TB of content returned with a video mime type, the actual size ingested, after de-duplication would be smaller. Similarly, we can deduce from the mime reports that we have crawled around 144TB of mp4 content, but the volume of de-duplicated mp4 files that we have actually ingested would be a smaller figure—one that we cannot currently pinpoint with precision.

Another issue with analyzing our content by mime-type is that webmasters are free to configure their web servers to return mime-types of their choosing. In fact, they can return altogether bogus mime-types such as "woof/woof", "xxx/xxx", "x-access/x-denied", etc. Consequently, the mime type reports contain inaccuracies, and thus cannot be relied upon for perfect analysis figures. The eventual hope is to be able to analyze the raw content within the compressed WARC/ARC files, which is stored in binary format, making it inaccessible without proper tools. One of the tools that we are currently exploring is the python based version of the "libwarc" software library. More information on this software project can be found at the links below:

http://netpreserve.org/projects/warc-tools-project
http://code.hanzoarchives.com/warc-tools

RELEVANT RECOMMENDED PRACTICES

Advice for File Archivists

- **Move video files off internal camera data storage, videotape, optical media or other unstable physical carriers to more stable storage media as soon as possible:** After each crawl is completed by the Internet Archive, we must use in-house tools to migrate the content to our local network, copy it to long term storage, process it for public access, and then copy it to public access storage.

Advice for File Creators and File Archivists

- **Document and use technical metadata:** We use BagIt metadata fields to store information on the storage containers (bags) that house the crawl content as well as separately generating a range of data reports for each crawl that provide more content-specific statistics.

- **Select video encoding and wrapper formats that are supported by downstream applications:** This principle applies to the Library's Web Archive in that our diversity of content must be stored in compressed WARC formatted files so that the Wayback Machine application can replay it.

When Following the Recommended Practices Is Not Practical

The project goals required some compromise and preclude following several Recommended Practices including

- **Select High Definition (HD) video encoding over Standard Definition (SD):** Currently we have opted to collect the high definition version of nominated YouTube videos. So technically, this principle does currently apply to our collection. However, given the volume of video content that we collect and the storage/processing limitations that we face, there is an argument to be made for why a sizeable Web Archive would specifically chose not to pursue HD quality when an SD option is available.

- **Develop selection criteria based on business needs to inform decisions on what files and/or formats to keep, especially if the same content is submitted in multiple video files:** As our case history demonstrates, the practice of following strict and comprehensive format and file quality selection is generally counterintuitive given the context of web archiving. Selection criteria focuses for web archiving activities focuses on content, and the nature of the internet implies that we will encounter, and must embrace, a broad diversity of formats.

- **Avoid multiple compression and decompression steps:** Given the dependence on the Wayback Machine to replay the archived content, it is necessary that to store that content in compressed WARC files. This provides the added benefit of reducing the total volume of our bulk storage, while supporting stewardship and the standardization methods of the greater Web Archiving community.

- **Stay within the same codec family if the video data is transcoded:** The Web Archiving project stores all content in compressed WARC files regardless of its source format.

LESSONS LEARNED

Deduplication

In coping with our expanding crawl outputs and the storage/processing challenges that they present, we have embraced the efficiency practice of deduplication. This process is particularly important in the context of crawling YouTube. While the content of a YouTube video page might change—with the addition of new comments, for example—the video element itself often remains the same. By subjecting our crawl outputs to deduplication, we are capable of capturing changes made to the site without burdening our storage systems and transfer procedures further by re-processing previously acquired content.

Adaptation

Our ongoing collection of YouTube content requires us to remain adaptable to the frequent structural changes to YouTube's backend video service structure. And while this goal of adaptability could serve as a guiding principle for the greater web archiving community, given the constantly evolving state of the internet, it emphasizes the importance of cultivating and sustaining technical and procedural standardization in areas that do remain under our control.

NATIONAL ARCHIVES AND RECORDS ADMINISTRATION
MIXING IT UP: WORKING WITH HETEROGENEOUS SETS OF FILE TYPES

CASE HISTORY SUMMARY

This case history will examine how to handle a group of heterogeneous files supplied on optical discs and/or a hard drive. Readers will become familiar with how to extract technical metadata from the files in order to gauge quality and identify salient characteristics. Tools for playback and transcoding a wide variety of files will be discussed along with some of the thought process that is used to determine which format (Preservation Master, Reproduction Master or Distribution Copy) should be created for long-term storage and preservation. A brief discussion of how to handle complex files that are stored in a directory structure is also included.

CASE HISTORY AUTHORS

Courtney Egan (courtney.egan@nara.gov), Audio Visual Preservation Specialist

John Powell (john.powell@nara.gov), Archivist

INTRODUCTORY INFORMATION

Institutional Background

The Audio-Video Preservation Lab at NARA[24] works with both audio and video materials and can handle legacy as well as modern formats. Six full-time employees and a supervisor staff the Lab and are responsible for reformatting records for preservation and access purposes. Equipment in the Lab includes several purpose-built reformatting workstations, a robotic videotape transfer system, automated quality control and transcoding tools and a set of applications for embedding and extracting technical metadata. The Lab may work on material from the overall NARA collection or from the Presidential Libraries and Regional NARA Offices around the country. Lab staff can also serve as subject matter experts for reformatting issues specific to audio-video recordings and media.

Collection Background

NARA is responsible for preserving the records of all federal agencies and is estimated to hold about 400,000 audio-video items from various parts of the U.S. federal government. Audio records range from disc recordings that date to the 1930-40s up to very recent submissions delivered as mp3 files on Flash cards; ¼" tape is probably the most prevalent format in the audio collection. Video records also span a large time frame and include significant amounts of 1-inch, U-matic and Digital Betacam tapes. It is only recently that NARA has begun to receive born digital records. The first example of born digital content arriving in the Lab was the Supreme Court audio recordings from the early 2000s; a small box of flash cards filled with that year's data was delivered to the Lab for quality control and preservation work. Other born digital projects have included NOAA field recordings on DVCPRO tapes; these were reformatted to AVI files with the original essence format (DVCPRO) maintained.

CASE HISTORY DETAILS

The archival unit that accessions electronic media at NARA received a large set of files from the Base Realignment and Closure Commissions from 2005. Many of these records were common file types and standard procedures were in place to process and transform them (if necessary). Mixed in with these common file types were several DVDs and CDs that contained audio-video content. Shortly after receiving this collection, the Electronic Records archivists took the precautionary step of moving the files from optical media formats to a hard drive. Beyond that they were unsure of what steps should be taken to preserve these items and they were having trouble even viewing some of the files so that they could review the content and determine whether or not it should become a permanent record. They sought

[24] http://www.archives.gov/

assistance from the Motion Picture and Sound archivists who suggested a versatile playback tool that they were familiar with and who also directed them to the Audio-Video Lab.

After using playback tools to determine that these records should be accessioned with the collection, one of the Electronic Records archivists worked with the AV Lab to determine next steps. Together lab staff and the archivists determined three requirements for what became a small pilot project: 1. Identify the technical characteristics of these files and supply this information to the archivists 2. Determine an intermediate format that these files should be transcoded to and 3. Migrate these to the intermediate format and deliver new copies to archivists. In order to identify and evaluate file characteristics for this heterogeneous group of material, the AV Lab needed a file analysis tool that supported a wide range of file types. MediaInfo worked well and had the capability export customizable sets of metadata. Once the files had been analyzed and technical characteristics exported to a spreadsheet, Lab staff could get a general impression of file size and quality. As mentioned above, many of these items had been stored on DVD and were in fact in the format of authored DVDs. Other file formats included the following wrappers: AVI (*.avi), Flash Video (*.flv), QuickTime (*.mov), MPEG-1 Video (*.mpg), MPEG-2 Video (*.mpg), ShockWave (*.swf) and Windows Media (*.wmv). Codecs used in file essences included Sorenson H263, AVID JFIF, Cinepack, Indeo4, MPEG-1 Video, MPEG-2 Video, MS Video, WMV3 for video and AC3, LPCM (Big Endian), MPEG-1 Audio Layer 2, MPEG-1L3, PCM and WMA2 for audio. The audio and video material ranged in quality from an overall bitrate of 1.48 Mbps (Megabits per second) to 9.33 Mbps, but no uncompressed or lossless compression codecs were used.

Because of the relatively low quality of the source material Lab staff determined that these items did not need to be migrated to our Preservation Master format, uncompressed video in an AVI wrapper. It was sufficient to migrate these items to our Reproduction Master format: MPEG-2 at 50Mbps. This is a stable and well-supported encoding of medium quality that generally does not introduce visual artifacts. For the items that originated as authored DVDs the Lab determined that generating an ISO file for each would be the most appropriate transformation activity. An ISO file allowed us to preserve menu functionality as well as the audio and video content; it also simplified file management because it transformed the set of several files that make up an authored DVD into a single file.

The last challenge in this case history was to identify a tool, or tools, that could transform the range of formats present. After some testing we determined that WinFF, the freely available GUI-based version of FFmpeg, met our needs. It was able to transcode to our exact specifications and it supported all of the source formats present. It was somewhat difficult to find a tool that could support the two flavors of Flash in our source files: *.swf and *.flv. Our commercial off the shelf transcoder was not able to work with these formats, nor was another piece of free software we had identified. WinFF proved to be the most versatile of the transcoding tools that we could identify; it has the robust capability of the FFmpeg code that it's built upon and the ease of use that a GUI brings with it. To transform the authored DVDs into ISO files we used free software called ImgBurn; it was simple and straightforward and worked well for us. It can also generate file-level MD5 checksums which could be an advantage in some workflows.

The Lab completed their portion of the project by quality checking the newly transformed files against the originals in order to ensure that no unwanted changes were introduced. We also provided copies of the new files and delivered a spreadsheet containing key technical metadata for each source file in the set. When complete, we handed off these deliverables to the electronic records archivists and they deposited a copy in our institutional repository for long-term storage and preservation.

SUMMARY OF VIDEO AND AUDIO DATA TECHNICAL SPECIFICATIONS

1. Source File - Lowest Quality

Video Data

- Container/wrapper: Windows Media
- Target total bitrate: 1.48 Mbps
- Timecode: none
- Frame size: 640x480
- Video codec: WMV3
- Video frame rate: 29.97
- Color encoding: unknown
- Chroma format: unknown
- Interlaced/Progressive scan: Progressive

Audio Data

- Audio channels: 2
- Audio codec: WMA2
- Audio sample rate: 44.1 kHz
- Audio bit depth: 16-bit

2. Source File - Highest Quality

Video Data

- Container/wrapper: MPEG-2
- Target total bitrate: 9.33 Mbps
- Timecode: none
- Frame size: 704 x 480
- Video codec: MPEG-2 Video
- Video frame rate: 29.97
- Color encoding: YUV
- Chroma format: 4:2:0
- Interlaced/Progressive scan: Interlaced, Top Field First

Audio Data

- Audio channels: 2
- Audio codec: AC3
- Audio sample rate: 48 kHz
- Audio bit depth: 16-bit

3. Normalized MPEG-2 File for Access Copies (non-authored DVDs)

Video Data

- Container/wrapper: MPEG-2
- Target total bitrate: 50.4 Mbps
- Timecode: none
- Frame size: 720x480
- Aspect ratio: 4:3
- Video codec: MPEG-2 Video
- Compression type: Lossy
- Video frame rate: 29.97
- Color encoding: YUV
- Chroma format: 4:2:2
- Interlaced, Bottom Field First

Audio Data

- Audio channels: 2
- Audio codec: MPEG Audio
- Audio sample rate: 48 kHz

4. Normalized File ISO file for Access Copies (authored DVDs)

Video Data

- Container/wrapper: MPEG 1/2
- Target total bitrate: ~4.9 Mbps
- Timecode: none
- Frame size: 720x480
- Aspect ratio: 4:3
- Video codec: MPEG 1 / 2

- Compression type: Lossy
- Video frame rate: 29.97
- Color encoding: YUV
- Chroma format: 4:2:0

Audio Data

- Audio channels: 2
- Audio codec: AC3
- Audio sample rate: 48kHz

RELEVANT RECOMMENDED PRACTICES

Advice for File Archivists

- **Determine criteria for when (if ever) it is appropriate to change the video file's technical properties (including normalization):** NARA-BRCC Case History project details the rationale for normalizing file types.

- **Retain the original video file as submitted if transcoding, normalizing or otherwise changing the video stream to meet business needs**: Originals are retained as submitted.

- **Select appropriate technical characteristics for the video encoding if transcoding, normalizing or otherwise changing the video stream to meet business needs:** NARA-BRCC Case History project details the rationale for selecting file characteristics for normalizing ingested files.

- **Move video files off internal camera data storage, videotape, optical media or other unstable physical carriers to more stable storage media as soon as possible:** Transfer video off optical discs and external hard drives into managed storage as soon as it can after accession of the collection.

Advice for Files Creators and File Archivists

- **Avoid multiple compressions and decompressions steps:** NARA-BRCC uses single-step transformation to move from the source files directly to the normalized intermediate formats.

- **Stay within the same codec family if the video data is transcoded:** Higher quality MPEG-2 source files remain in MPEG-2 for the target format, MPEG-2 at 50Mbps.

- **Select video encoding and wrapper formats that are well-supported now and future focused:** MPEG-2 @ 50Mbps is a common implementation of the MPEG-2 standard and should be supported most applications.

- **Select video encoding and wrapper formats that are non-proprietary:** MPEG-2 is standardized through the ISO 13818 document set.

- **Select video encoding and wrapper formats that are supported by downstream applications:** MPEG-2 @ 50Mbps is a common implementation of the MPEG-2 standard and should be supported most applications.

- **Select video formats that are standardized and well-documented:** Both MPEG-1 and MPEG-2 are well established and well documented through ISO.

When Following the Recommended Practices Is Not Practical

The project goals required some compromise and preclude following several Recommended Practices most notably in the choice of lossy compression:

- **Select uncompressed video encoding over compressed encoding:** This case history project did not select an uncompressed target format because the source material already was highly compressed. In addition, data storage was limited so the increased file size would be problematic.

- **If compression is used, select mathematically lossless compression over visually lossless or lossy compression:** The case history uses MPEG-2, a visually lossy compression, because lossless compression would have resulted in larger file sizes that weren't justified by the source material.

LESSONS LEARNED

One of the first lessons we learned in this project was that it can be difficult to work with a diverse set of file formats so an important first step is to identify the technical characteristics of the files. You could look at the highest and lowest quality files in order to give you a sense of the range of quality in the collection and then make decisions based on that information. Another trend we picked up on was that open source and freely available tools may be more likely to support older and more obscure file formats than commercial products would be. Lastly we found that it is important to compare your source files with the newly generated files to ensure that no artifacting has been introduced.

The archivists that we worked with had also taken a couple of helpful steps shortly after these records were received in their unit. We found it very useful that the archivists we were working with had moved these files off of optical disc and onto a hard drive shortly after receiving them. This meant that our failure rate was considerably lower; we only had one file (out of about 40 items) that was not able to be transferred successfully from disc. Additionally, the archivists had done the important pre-work of organizing the source files into directories. This meant that each item had an identification number and that complex files were grouped together under that identification number.

SMITHSONIAN INSTITUTION ARCHIVES
PRESERVING CONTENT FROM AUTHORED VIDEO DVDS

CASE HISTORY SUMMARY

During the past three years there has been tremendous growth in collections that contain digital video at the Smithsonian Institution Archives (SIA).[25] In 2013 SIA decided to work with its collection of Smithsonian Channel programming on authored video DVDs as a pilot. SIA is not alone in the challenges of authored video DVD contents and believes both file creators and archivists can learn from our experiences. Our best practices are to ingest a copy of digital source files from the media upon receipt to create preservation masters and access copies whenever feasible. This applies to all formats (text, images, database, audio, etc.) and media. Since DVDs are optical discs the contents need to be ingested onto backed-up networked servers, external drives, and LTO tapes to ensure copies can be made and to undergo preservation processes. Simply copying the files (VOB, IFO, and BUP) directly off the authored video DVD breaks the menu functionality that one sees when a DVD is launched from a player or computer. SIA sought a solution that would retain as much as possible from the original DVD.

CASE HISTORY AUTHORS

Lynda Schmitz Fuhrig (schmitzfuhrigl@si.edu), Smithsonian Institution Archives

INTRODUCTORY INFORMATION

Institutional Background

SIA captures, preserves, and makes available to the public the history of an extraordinary institution. From its inception in 1846 to the present, the records of the history of the Institution -- its people, its programs, its research, and its stories -- have been gathered, organized, and disseminated so that everyone can learn about the Smithsonian. The history of the Smithsonian is a vital part of American history, of scientific exploration, and of international cultural understanding.

SIA has about 38,000 cubic feet of collections and 6 TB of born digital collections. The Archives has almost 700 collections that include born digital materials dating back to the 1980s. Types of files include text, images, databases, audio, video, email accounts, websites, and social media accounts.

There is a full-time staff of 25 at SIA, and the Digital Services Division (DSD) has five employees. DSD has one electronic records archivist who handles all born digital accessions and is assisted by interns and volunteers. A digitization and imaging specialist also is on staff.

Collection Background

The Smithsonian Institution Archives continues to see growth among digital video that is created for and by the Smithsonian Institution's many museums, research facilities, and offices. Four collections contain 546 authored video DVDs from Smithsonian Channel[26] programming, and it is the responsibility of SIA to ensure the contents remain playable and accessible for the long term (Smithsonian Channel is owned by Smithsonian Networks, a joint venture between the Smithsonian and Showtime.). As part of Smithsonian Network's contract with the Smithsonian Institution, channel programming is transferred to SIA for permanent accession at regular intervals. Smithsonian Networks also has its own archive of the programming that comes off tape and file-based media.

These collections from 2011-2013 were burned to DVDs by Smithsonian Networks staff and then transferred to SIA. As authored video DVDs, the video and audio are limited in quality due to having lossy compression, a small frame size, and lower bit rates.

---

[25] http://siarchives.si.edu/
[26] http://www.smithsonianchannel.com/sc/web/home

The lifespan of DVDs is considered short, with expectancy figures listed anywhere from two years up to 25 years.[27] Many factors play into the longevity including the manufacturing process and handling and storage of the media. These reasons, along with possible software and hardware challenges, make it important to copy the DVD contents as soon as possible. Another issue is that computer manufacturers continue to phase out optical disc drives.

CASE HISTORY DETAILS

With this pilot we wanted to test software to see what could give us a midterm access copy that would at least retain some of the menu functionality if possible. In 2011 we made the decision to create ISO files[28] of authored video DVDs as the preservation master so we would have a bit-for-bit version of the contents and be able to play back the video without the media. The ISOs are created in batch with a Ripstation, which can process up to 80 DVDs at a time. The Smithsonian Channel programming was a good candidate for a summer internship in 2013 because of the number of DVDs and that it was all one format.

Research included:

- Testing software to create MPEG-2 files from the DVD contents.
- Maintaining menus (we were only able to capture the initial screen with no functionality).
- Creating metadata as XMP sidecars.
- Creating workflows and importing contents to the Smithsonian's enterprise DAMS using BagIt.

The Process

Copying the files (VOB, IFO, and BUP) directly off the DVD breaks the menu functionality that one sees when a DVD is launched from a player or computer. SIA's solution has been to create an ISO of the DVD. This ISO file can be mounted to a computer for viewing with appropriate player software as if it was an actual DVD with the user menus still operational. This serves as the preservation master.

An access copy also has been created by stringing together the individual VOB files and using FFmpeg software to create a single MPEG-2 with an MPEG wrapper. While the access MPEG-2 file lacks the menu's functionality, there is a brief screen of the menu at the beginning of playback, which had been lost when using other software applications. Either a copy of the ISO or MPEG-2 can be provided for playback.

This work also has guided procedures with authored video DVDs received by SIA that are not Smithsonian Channel programming.

SUMMARY OF VIDEO AND AUDIO DATA TECHNICAL SPECIFICATIONS

Video Data

- Container/wrapper: MPEG-2
- Target total bitrate: Varies – 9500 Kpbs is maximum
- Timecode: Varies
- Frame size: 720x480
- Aspect ratio: 4:3
- Video codec: MPEG-2
- Video frame rate: 29.97 fps

[27] http://www.archives.gov/records-mgmt/initiatives/temp-opmedia-faq.html
[28] http://www.digitalpreservation.gov/formats/fdd/fdd000348.shtml

Audio Data

- Audio channels: 2
- Audio codec: MPEG version 1, Layer 2
- Audio sample rate: 48.0 kHz

RELEVANT RECOMMENDED PRACTICES

The project goals permitted following many of the Recommended Practices including:

Advice for File Archivists

- **Document the original order, especially camera-created file structures:** SIA-DVD follows the VOB order when creating a MPEG-2 access file.

- **Identify the file characteristics at the most granular level possible, including the wrapper and video stream encoding**: MediaInfo is used on the contents in this case history, as well as used to review the MPEG-2 files that are created.

- **Determine criteria for when (if ever) it is appropriate to change the video file's technical properties (including normalization):** Video files on DVD are not a long-term option for preservation or access.

- **Retain the original video file as submitted if transcoding, normalizing or otherwise changing the video stream to meet business needs:** DVD originals are retained as submitted as part of best practices.

- **Select appropriate technical characteristics for the video encoding if transcoding, normalizing or otherwise changing the video stream to meet business needs:** Case history details the rationale for selecting file characteristics for normalizing ingested files.

- **Retain original timecode(s) if provided, even if you generate a new high integrity continuous master timecode:** If timecode is present, it is retained.

- **Retain original frame rates if the video data is transcoded, even when they are beyond the standard 24 - 30 fps:** Standard 29.97 fps is retained.

- **Move video files off internal camera data storage, videotape, optical media or other unstable physical carriers to more stable storage media as soon as possible:** SIA transfers video off DVD as soon as it can after accession of the collection.

Advice for File Creators and File Archivists

- **Document and use technical metadata:** Metadata is noted on spreadsheets and other documentation is created.

- **Select video encoding and wrapper formats that are well-supported now and future focused:** SIA-DVD uses MPEG-2, which should work with many video players including WMP, VLC, and QuickTime. ISOs at this time can be mounted on computers for playback. With an ISO copy, SIA can revisit the file for additional processing as software is developed to possibly create a "more accurate" access file, such as better audio conversion.

- **Select video encoding and wrapper formats that are non-proprietary:** SIA-DVD uses MPEG-2, standardized through the ISO 13818 document set.

- **Select video encoding and wrapper formats that are supported by downstream applications:** File remains MPEG-2. VOB is based on the MPEG program stream.

- **Select video formats that are standardized and well-documented:** Both MPEG-1 and MPEG-2 are well established and well documented through ISO.

- **Select video formats with greater capacity for embedded metadata over less metadata capacity**: MPEG-2 allows for an XMP file, which lives with the corresponding file and is also imported into the DAMS.

When Following the Recommended Practices Is Not Practical

The project goals required some compromise and preclude following several Recommended Practices most notably in the choice of lossy compression:

- **Select uncompressed over compressed.** Authored video DVDs contain video that is lossy and compressed. Trying to convert to something higher can result in problematic files that either have artifacts, poor syncing, or other issues. The course of action is to preserve the exact specifications as much possible.

- **If compression is used, select mathematically lossless compression over visually lossless or lossy compression:** The case history uses MPEG-2, a visually lossy compression, because lossless compression would have resulted in larger file sizes that weren't justified by the source material.

- **Avoid multiple compression and decompression steps:** The video content on the authored DVDs is lossy and compressed and is normalized to MPEG-2 to meet business needs.

LESSONS LEARNED

Software Limitations

At first SIA was going to create MPEG-4 access files because the Smithsonian Institution Libraries was doing this in a DAMS-related project, but the files SIA created would not play back within the Smithsonian's DAMS. MPEG-2 was adopted at that point. Various software applications that were tested also either resulted in artifacts on output, timecode issues, missing programming, and/or no menu page. Nevertheless, that is the point of testing software within a pilot. Those files with the timecode issues would sometimes play back without issue though.

Working with the Smithsonian's DAMS team in the Office of the Chief Information Officer (OCIO), SIA started testing FFmpeg, a well-established and free command-line tool for converting, streaming, and recording video and audio. SIA successfully created one VOB by stitching all the VOB files together and then used FFmpeg to transform that VOB file into a playable MPEG-2 with an MPEG wrapper that is supported within the enterprise DAMS. FFmpeg also retains original timecode of the authored video DVD from the concatenated VOB files, in addition to any original subtitles on the disc and the menu page.

During our work with these collections SIA used two different scripts to create the MPEG-2s, in order to have satisfactory files within the DAMS that create and play back a proxy through the DAMS transcoders. The first script (ffmpeg -i [input file.vob] -c:v copy [output file.mpg]) retained the source encoding for the video but re-encoded the audio. Unfortunately, the majority of these files failed within the DAMS before and after a software upgrade. The second script used in 2014 (ffmpeg -i [input file.vob] -vcodec mpeg2video -pix_fmt yuv420p -me_method epzs -threads 4 -r 29.97 -g 15 -s 720x480 -aspect 4:3 -b 9500k -bt 400k -acodec mp2 -ac 2 -ab 192k -ar 48000 -async 1 -y -f mpeg [output file.mpg] re-encodes both the video and audio with specs as close as possible to the source. Note: This second script is a tweak that George Blood Audio Video Film documented in a report for the Library of Congress about extracting data from authored DVDs.[29] These files successfully displayed and played back in the Smithsonian's DAMS. It should be noted that the MPEG-2s created with the first script did play back in video players outside of the DAMS

Workflow Tweaks

The DVD contents were being saved to an external drive for processing work. Once ISOs and MPEG-2s were created, standard naming applied, and metadata added, Bag-It was used to transfer the files to the server for import

---

[29] Preserving Write-Once DVDs: Producing Disk Images, Extracting Content, and Addressing Flaws and Errors
By George Blood Audio Video Film for the Library of Congress
http://www.digitizationguidelines.gov/audio-visual/documents/Preserve_DVDs_BloodReport_20140901.pdf

into the DAMS in order to verify integrity of the files. It was decided the BagIt step was not needed since checksums are generated of the ISOs and the MPEG-2s and upon ingest into the DAMS for comparison.

One Size Does Not Fit All

This project did not provide solutions for all AV contents SIA encounters on DVD. This case study does not address video on DVD that contains standalone MOV, AVI, MPG, and SWF files. SIA also has not been able to apply the workflow successfully to DVDs created in a Mac environment to date. Research done in-house in 2011 reviewed digital video assets within SIA's collections at that time to determine what was playable with current software and hardware, providing a baseline document.

What Is Acceptable

SIA can accept the transcoding of audio to MPEG-1 with the access copy. FFmpeg's settings create this because it appears that the MPEG wrapper has issues with AC3.

Future Files

There has been an initial conversation with the Smithsonian Channel archivist about a possible pilot of having the files directly transferred to SIA or the DAMS and completely avoiding the DVD step in the future. This will mean better quality video and fewer processing steps.

While it is desired to be able to be part of the digital file lifecycle from the beginning, oftentimes cultural heritage institutions receive materials after their creation, which means they are stuck with what they get. This case study provides one approach these organizations should be able to adopt.

FOR MORE INFORMATION

See two posts from the SIA Blog, *The Bigger Picture*

- **What Are You Watching? (**http://www.siarchives.si.edu/blog/what-are-you-watching)
- **And Action: The Ins and Outs of DVD Video Preservation** (http://www.siarchives.si.edu/blog/and-action-ins-and-outs-dvd-video-preservation)

SMITHSONIAN INSTITUTION
INGEST AND ARCHIVING OF CAMERA ORIGINAL VIDEO FILES WITH AN ENTERPRISE DAMS: THE RECOVERING VOICES COLLECTION, NMNH, DEPARTMENT OF ANTHROPOLOGY

CASE HISTORY SUMMARY

Camera original video files are often delivered off-camera in nested folder structures. Working with an Enterprise Digital Asset Management System that supports the ingest of files versus nested folder structures without developing a custom ingest presents challenges for the processing and archiving of nested and hierarchical camera original born digital video files. The challenges encompass maintaining original order, harvesting relevant metadata, analyzing risks in file technical specs, verifying file authenticity, and developing sustainable workflows. This case study will illustrate the workflow adopted from receipt of files from content creator to the Smithsonian DAMS repository for archiving and reuse.

CASE HISTORY AUTHORS

Crystal Sanchez (sanchezca@si.edu), Video and Digital Preservation Specialist, DAMS - OCIO, Smithsonian Institution

Stephanie Christensen (christensens@si.edu), IT Specialist, DAMS - OCIO, formerly NAA Digital Imaging Manager, Smithsonian Institution

Isabel Meyer (meyeri@si.edu), DAMS Manager - OCIO, Smithsonian Institution

INTRODUCTORY INFORMATION

Institutional Background

Founded in 1846, the Smithsonian is the world's largest museum and research complex, consisting of 19 museums and galleries, the National Zoological Park, and 9 research facilities. One of the units, the Smithsonian's National Museum of Natural History[30] (NMNH) serves as one of the world's great repositories of scientific and cultural heritage through its research, collections, education, and exhibition programs. NMNH's Department of Anthropology[31] houses the Recovering Voices Initiative[32] (RV), which is dedicated to nurturing efforts to document and revitalize endangered languages and knowledge systems through research, collaboration, and resources.

At the Smithsonian Institution, stewarding digital assets – managing how they are collected, stored, preserved, secured, accessed, and exhibited – is an institution-wide concern. The Institution has deployed an enterprise Digital Asset Management System (DAMS), managed by the Office of the Chief Information Officer[33] (OCIO), a central IT department, to manage this effort. The SI DAMS currently holds over 5 million images, audio, video, and supporting documentation records. The system consists of software, hardware, and database resources coupled with management tasks and decisions surrounding the ingest, annotation, cataloging, access, storage, retrieval, distribution, and preservation of digital assets. It provides a place for assets that may be stored in multiple formats and locations, while providing centralized and cost-effective preservation, security, backup, and recovery.

Utilizing a vendor supported browser based application, each unit and/or unit department within the Smithsonian Institution works with the SI DAMS team to build a foundation according to their specific needs. This approach

[30] https://www.mnh.si.edu/
[31] http://anthropology.si.edu/
[32] http://anthropology.si.edu/recovering_voices/index.htm
[33] http://www.si.edu/ocio/

provides a solution where individual units can organize and manage their own collections and maintain their unique areas of expertise, all within the framework of an enterprise-wide service deployed and maintained by OCIO. The DAMS application, available to all Smithsonian employees, offers specific features that will be used as part of the workflow solution in this case study.

Collection Background

The Recovering Voices Initiative creates an array of digital assets as part of their work to document endangered languages and knowledge systems. These assets include image, audio, and video files.
The video files are currently delivered as camera original files by the Recovering Voices production staff to the National Anthropological Archives[34] (NAA) for processing and archiving. The files are delivered on hard drives with default naming conventions in a nested folder structure created by the camera and with little or no metadata. The National Anthropological Archives advises the RV team[35] on file naming, appropriate metadata, and file formats for sustainability. The SI DAMS archives the assets but its ingest capability is limited to the file level, not a structured folder level. The files are processed and ingested to the SI DAMS by NAA staff in partnership with the staff at Recovering Voices.

CASE HISTORY DETAILS

Before ingest to the SI DAMS, the DAMS staff documented the following requirements which needed to be accommodated in an ingest workflow for these video collections:

- The Recovering Voices staff explained that their most pressing needs as content producers are the general back-up of files, storage, search, and retrieval. There is limited to no post-production work done between creation and archive.
- Upon delivery to the archives, files are organized by collection/project/shoot name. Camera original files are delivered to the archive as direct copies from camera to hard drive in their original hierarchical folder structures with camera generated numbers and letters.
- NAA Digital Imaging Manager expressed the desire to create a hierarchical folder archival structure in DAMS based on project/shoot, and review file naming protocols and metadata for eventual establishment of file naming conventions and basic metadata.

With these requirements in mind, the DAMS staff outlined an ingest workflow which includes the following facets:

- **Determine what is received** by compiling an inventory of files received with technical specifications of the file characteristics.
- **Determine any risks** including format obsolescence, relationships among and between files, software dependencies, etc.
- **Determine end result of files to be archived**: What is the short and long term goal from producers, archivists, unit, and the public?
- **Determine the criteria for selection and retention** when content is presented in multiple file formats.
- **Determine the impact of original order and interfile relationships** and create a plan for documenting the original relationships.

---

[34] http://www.nmnh.si.edu/naa/

[35] For the initial case study, the materials were received by the archive as is, without consultation with archive. After the initial review of files and as the RV team moves forward with other projects, file naming, appropriate metadata, and file formats for sustainability are being recommended to the RV team so that some of the work can be done at initial capture and not in post- production. The goal is to provide RV with guidelines for "best practices" in capture that can be integrated into their documentation workflow.

- **Establish metadata requirements:** Analyze any embedded or provided metadata. Determine what metadata should be retained and create minimal level metadata recommendations for archiving.

- As part of RV Digital Asset Management Plan, **establish file naming guidelines** and DAMS Folder Structure.

- **Create realistic project management plans** including required resources and roles and responsibilities as well as timelines for archiving files and a guide for future timeline goals between producer and archive.

- **Create detailed workflow documentation** including a step by step guide detailing the process to move assets and associated data from producer to archive to DAMS ingest.

SUMMARY OF VIDEO AND AUDIO DATA TECHNICAL SPECIFICATIONS

Master Camera Original Video Data

Analysis of the files and accompanying XML files indicate that the files from the Recovering Voices project were created with a Canon XF100 camera, which shoots MPEG-2 wrapped MXF files at up to 50 mbps with 4:2:2 color space. The master files have the following technical specifications:

- Video stream: Canon HD422 MPEG2
- Audio stream: PCM
- Wrapper: MXF OP1a
- Color space: YUV 4:2:2
- Bit rate: 50Mbps variable, interlaced
- Resolution: 1920x1080i
- Aspect ratio: 16:9
- Frame Rate: 29.97

Other Video Files

The MXF files are accompanied by derivative files: QuickTime (*.mov) and Flash (*.f4v) video files, *.xml files, *.cif files and *.sif files. The *.cif files are Common Intermediate Format (CIF) or small proxy files that the camera makes for viewing, probably to allow for logging/monitoring on site through the camera. The *.sif files are Source Input Format (SIF) files that exist for directional purposes for import to an editing program. There is concern with XDCAM-like files in that if the original folder structure is stripped, there could be some loss by an editing suite in reading the original output. Retention of these files really depends on the use of this material before archiving and use in the future. Producers were consulted and it was determined that they do not edit the files before archiving, and that these files were not necessary to retain.

RELEVANT RECOMMENDED PRACTICES

The project goals permitted following a wide variety of the Recommended Practices including:

Advice for File Archivists

- **Document the original order, especially camera-created file structures:** NAA archivists using the SI-DAMS have created a folder structure to replace (not entirely replicate) the original file directory structure from within the SI-DAMS interface and files will be placed in this structure for easy retrieval and to retain original order. File names are changed to conform to the established file naming conventions. The original file names (which reference original order folder structure) are included in the metadata for each file.

- **Document relationships between the video object and other files, such as closed captions, scripts, location notes and other supplemental material:** The SI DAMS, through its application, introduces a ‘folder’ organizational functionality. NAA archivists have created a folder structure to replace (not entirely replicate) the original file directory structure from within the SI DAMS interface and files will be placed in this structure for easy retrieval and to retain original order.

- **Identify the file characteristics at the most granular level possible, including the wrapper and video stream encoding:** Ingest tools that include MediaInfo, Exiftool, and the DAMS Transcoders read and extract technical information from the file, including MIME type, and map this information to DAMS metadata fields to allow for the validation, capture, and documentation of key technical file level information. This data includes audio and video stream encodings.

- **Develop selection criteria based on business needs to inform decisions on what files and/or formats to keep, especially if the same content is submitted in multiple video files:** File packages were inspected and research into the production path was done. The MXF video files were determined to be the original, master files, and they will be retained. The MOV files are derivatives of the master and will also be retained for future use and easy retrieval by the producers. The F4V (flash) files are derivative at-risk files. The XML file was found to not contain any information other than the camera specs, which will be included in the metadata. After speaking with the producers, the CIF and SIF files were found to be not essential or useful for archiving. The F4V, XML, CIF, and SIF files will be discarded, while the MXF and MOV files will be retained when present.

- **Move video files off internal camera data storage, videotape, optical media or other unstable physical carriers to more stable storage media as soon as possible:** Files are delivered from the producers to the archive on external hard drives. Files are moved from the hard drives to the DAMS staging servers as soon as received and held there until fully processed and ingested to the repository.

Advice for File Creators and File Archivists

- **Use XML-based metadata schemas with strong support for digital video:** Mapping to PREMIS or other common schemas can easily be done with the SI-DAMS standard XML format for each asset.

- **Document and use technical metadata:** SI-DAMS uses MediaInfo and Exiftool to read the technical information from the files at the point of ingest, including limited supported descriptive fields (XMP). The DAMS ingest transcoders create proxy video files for viewing in the DAMS, at which point technical information (video and audio codecs, bit rate, frame rate, frame height and width, aspect ratio, audio sample rate, number of audio tracks) is generated. This makes them reportable for obsolescence monitoring. More fields are recommended for manual data entry: originating format, coding history, color space, capture device, caption format.

- **Determine criteria for when (if ever) it is appropriate to change the video file's technical properties (including normalization):** Upon ingest to the SI DAMS, a preview copy of each file is made with the DAMS transcoders. Only major formats are supported, allowing us to target and troubleshoot proprietary and/or obsolete formats, and then make decisions about moving forward with them. Files in this case study are supported files and can be ingested to the DAMS with preview files made. We do not normalize and we do not currently migrate at the point of acquisition/ingest.

LESSONS LEARNED

Having a DAMS centralized application with a browser-based user interface for access to the digital repository allows for varied, diverse collections across Smithsonian units to be managed by the collection holders while also providing the support of a centralized repository. Utilizing one repository also allows for the adoption of consistent metadata policies and centralized auditing and preservation policies.

The ingest process for the SI DAMS, including the integrated automated processes, aligns with many of the FADGI Recommended Practices for Archiving Born Digital Video. Automating many of the file identification, authenticity, and technical and administrative metadata needs, among others, makes this a sustainable practice.

For the archivists processing the collections, working with the content creators is not always an easy process. Often decisions that are best for the collection are made independent of any in-depth consultation with the content creators, due to the size of the collection and the expedient need to process it to mitigate technological risk. Having a DAMS application allows for at-risk media to be secured and moved off of curators' local storage.

More direct work with content creators will help foment some of the strategies documented above, allow for a more seamless transition from producer to archive, and fill gaps in metadata that can only be provided by content creators. Using the SI DAMS allows for the initial ingest process of archiving with a loop back to content creators, who can later access their files through the application interface, enrich metadata by cataloging within the DAMS, and view, inspect and retrieve videos whenever they need them. This process also provides an efficient method to ensure protection of the files with replication and back-up, and allows collections staff to provide recommendations to the content creators as they move forward in production of materials, allowing for a more informed and sustainable workflow from creation to storage.

FILE CHARACTERISTIC COMPARISON TABLES

CREATING BORN DIGITAL VIDEO CASE HISTORY FILE SPECIFICATIONS SUMMARY TABLE

The file characteristics for the three *Creating* case histories are summarized in the table below.

- **LC-AFC-CRHP:** Library of Congress American Folklife Center Civil Rights History Project
- **NOAA-OkEx:** National Oceanic and Atmospheric Administration Okeanus Explorer
- **VOA-MMAM**: Voice of America Metadata for Media Asset Management

	LC-AFC-CRHP	**NOAA-OkEx**	**VOA-MMAM**
Video Data			
Container/wrapper	Apple QuickTime (.mov); MPEG-4 format	Apple QuickTime (.mov)	Apple QuickTime (.mov)
Target total bitrate	220 Mbps	145 Mbps	25 Mbps
Timecode	SMPTE	SMPTE (Control Clock set to UTC)	
Frame size	1920 x 1080i	1920 x 1080	720 x 480
Aspect ratio	16:9	16:9	4:3
Video codec	ProRes HQ (422)	ProRes (422)	DV25
Compression type	Lossy	Lossy	Lossy
Video frame rate	29.97 fps	29.97 fps	29.97 fps
Color encoding	YCbCr	YUV	YUV
Chroma format	4:2:2	4:2:2	4:1:1
Bit rate	10-bit		
Scan order	Interlaced	Interlaced scan (Top Field First)	Interlaced
Video frame rate mode	Variable	Constant	
Audio Data			
Audio channels	2	4	2
Audio codec	PCM	PCM	PCM
Audio sample rate	48 kHz	48 kHz	48 kHz
Audio bit depth	24 bit	24 bit	16 bit

ARCHIVING BORN DIGITAL VIDEO CASE HISTORY FILE SPECIFICATION SUMMARY TABLE

The file characteristics for the three *Archiving* case histories which utilize normalization are summarized in the table below.

- **LC-NAVCC-VEF:** Library of Congress Packard Campus of the National Audio-Visual Conservation Center Video Evergreen Format
- **NARA-BRCC:** National Archives and Records Administration Base Realignment and Closure Commissions project
- **SIA-DVD**: Smithsonian Institution Archives Authored DVD project

Two *Archiving* case histories do not change file characteristics for ingest so they are not included in the summary chart.

- **LC-WebArch-YouTube:** Library of Congress Web Archiving You Tube Harvesting
- **SI-DAMS:** Smithsonian Institution Digital Asset Management System

	LC-NAVCC-VEF	**NARA-BRCC**		**SIA-DVD**
File Purpose:	**Normalization for archiving**	**Normalization for access (non-authored DVDs)**	**Normalization for access (authored DVDs)**	**Normalization for access (authored DVDs)**
Video Data				
Container / wrapper	MXF OP1a	MPEG-2	MPEG 1/2	MPEG version 1, Layer 2
Target total bitrate	Average of 90 Mbps for SD; 800 Mbps HD (uncompressed HD input)	50.4 Mbps	~4.9 Mbps	Varies – 9500 Kpbs is maximum
Timecode	Native	None	None	Varies
Frame size	Native	720x480	720x480	720x480
Aspect ratio	Native 4:3 for SD; 16:9 for HD	4:3	4:3	4:3
Video codec	JPEG2000 Lossless reversible 5/3	MPEG-2 Video	MPEG 1/2	MPEG-2
Compression type	Lossless	Lossy	Lossy	Lossy
Video frame rate	Native	29.97	29.97	29.97
Color space	Native	YUV	YUV	
Chroma format	Native 4:2:2 for YPbPr, 4:4:4 for RGB	4:2:2	4:2:0	
Bit-depth	10-bit			
Scan Order	Native	Interlaced, Bottom Field First		
Audio Data				
Audio channels	Native	2	2	2
Audio codec	PCM	MPEG Audio	AC3	MPEG-1
Sample rate	48 kHz	48 kHz	48 kHz	48 kHz