DOI: 10.2352/issn.2168-3204.2016.1.0.117

The Evolution of Motion Picture Digitization at the National Library of Medicine

John P. Rees, John P. Doyle, Doron Shalvi, National Library of Medicine (USA)

Abstract

In September 2010, the National Library of Medicine (NLM) launched Digital Collections [1], a Fedora Commons-based repository which allows rich access to and the preservation of digital content important to the long-term stewardship of NLM collections. Included in this initial launch were 11 digitized motion pictures from NLM's History of Medicine Division (HMD). NLM's Digital Collections repository has evolved and grown, and now contains over 200 motion picture titles. This article examines the content modeling, selection, workflows, and software used to produce repository content, and presents our recent movement to digitization of motion picture content as a true preservation solution.

Background

NLM collects, preserves, and makes available motion pictures, films, and videos which are important to the research needs of scholars and historians in medicine and public health, as well as selected examples of audiovisual works produced for use by health professionals in continuing medical education, patient instruction, or health education. The Historic Audiovisual Program (HAV) relies primarily on donations to collect commercial and non-commercial content generally produced or created up to 1970. Much of this content is in the public domain due to its age or the transfer of rights trough deed of gift. NLM's general collection collects primarily commercial content created after 1970 with most of the content still under copyright protection. NLM has a long history of preserving and providing access to these collection materials via analog technologies, but without a staff of engineers, professional video and film experts, or studio equipment. Beginning in the 1980s, NLM routinely outsourced the preservation of audiovisual materials and copied original film to polyester 'safety' film and copied video to BetacamSP tape as mezzanine copy masters, with users accessing VHS or DVD copies; the original source is stored, untouched, in one of NLM's cold vaults in an Iron Mountain cave facility in Boyers, PA for long-term preservation. NLM has long held an outreach goal of providing online access to these films, but technology, policies, and resources were barriers to these efforts. In 2009, development work began in earnest to build a "digital repository", and public domain motion pictures from HAV's holdings were selected as one of the pilot formats.

With a clear mandate and additional resources there were still many hurdles to overcome. Most important among these was the Dept. of Health and Human Services' (HHS) strict Section 508 Accessibility policy covering online video: captions containing accurate and synchronized transcripts must accompany any online video content. Even without a stable of engineers and professional-grade equipment, digitizing video content for access soon became the easy part; creating transcripts and captions and providing access to the complete package was much more challenging. Fortunately our repository developers already started an experimental project to develop a video player with search capabilities for the NLM Director. We decided to incorporate this video player into our repository technology stack.

Content Selection

For our pilot test, we selected the 11 oldest cataloged motion pictures that were in the public domain. These were mainly World War II public health films produced by the US Navy and Army and likely to be the most difficult use cases. Films from the 1940s presented some quality challenges - audio fidelity is typically poor, and image resolution grainy. On the plus side, the content was copyright-free, and the films were some of our most engaging and widely requested. Who would not like to see a young, uncredited Gene Kelly playing the part of a shell-shocked sailor in *Combat Fatigue Irritability,* or learn good battlefield dental hygiene techniques in *Dental Health* [2]?

Curated content currently drives our selection criteria, rather than large-scale or mass digitization methodologies. However, as discussed below, we intend to introduce a more preservation-centric paradigm in 2016. Titles are drawn from existing subject guides or other thematic content and pass a review that considers content, quality, audience, and other restrictions. For example, an in-house DVD education module produced two early projects: *The Public Health Film Goes To War,* and *Tropical Disease Motion Pictures* (TDM). Two titles proposed for TDM were rejected on grounds laid out by our Access to Personally Identifiable Health Information policy.

Content Modeling

We began our modeling effort by identifying several use cases important to NLM:

- Digital Preservation of the highest quality digital file reasonably available;
- Native playback of video and captions over the open Web;
- Playback of video and/or captions on non-native devices, e.g. local desktop applications, handheld devices;
- Discovery of and rich search and interactivity within our videos;
- Export/download of master files and selected user access formats for external re-use; and
- Data Harvesting / Data Mining

DOI: 10.2352/issn.2168-3204.2016.1.0.117

Several work areas were broadly identified to satisfy these use cases, including:

- Digital Preservation
- Transcriptions and Captioning
- Broad Access and Derivative Creation
- Video Search

Each of these work areas is described in more detail in subsequent sections of this paper.

Our video content model was designed to satisfy these use cases independent of our choice of software platform and technologies. Working off some common assumptions made for a concomitant book digitization pilot project, this model originally read like an overwhelming laundry list of files:

Video files, including:

- **Master MPEG-2** (.mpg), native size, 44.1/48.0 Khz, >200 Kbps bit rate
- **Quicktime**, 640x480 pixels, derivative (.mov), 2.6 Mbps, 48 Khz audio
- **Windows Media**, 640x480, derivative (.wmv), 2.7 Mbps
- **H.264**, 640x480, derivative (.mp4) 1.25 Mbps, 44.1 Khz AAC audio
- **Video Player**, 480x360, H.264 derivative (.mp4) 375-575 Kbps, 22.050 Khz AAC audio
- **30 second clip** H.264 derivative (.mp4)
- **Iphone**, 480x360, H.264 derivative (.mp4), 1.0 Mbps

Caption files, including:

- Quicktime SMIL **caption** file (.smil/.txt)
- Quicktime SMIL **transcript** file (.txt)
- W3C **DFXP** caption file (.xml) [3]
- **Magpie** project caption master file (.magpie)
- Plain **text** transcript file (.txt)

Metadata files, including:

- **MARC XML** (derived from ILS; definitive descriptive metadata store)
- **DMDINDEX** (local XML descriptive metadata transformed from MARC XML, for repository SOLR index)
- **Dublin Core**

Still shots, including:

- 640x480 **large** poster image (.jpg)
- 320x240 **medium** poster image (.jpg)
- 160x120 small poster image (.jpeg)

Digital Preservation

From the outset, we have been hesitant to characterize our motion picture digitization project as "preservation oriented." It is more accurately described as an "access project with some preservation considerations." Regardless of how enticing it may sound to deliver this content solely over YouTube or social media outlets for widespread access, we also wanted to meet NLM's preservation mandate as well.

Best practices in film preservation still have not coalesced around any single preservation-worthy file format(s), contributing to the delay of more robust digitization at NLM. There are myriad choices for both codecs and file wrappers, each with their own set of pros and cons. Paralysis can often be the result. NLM looked to existing standards and like-minded implementers for solutions. We analyzed several best practice recommendation publications, such as the U.S. National Archives and Records Administration's (NARA) Reformatting Approaches website [4] and the Library of Congress's Sustainability of Digital Formats website [5], learning from our colleagues at these two sister institutions through participation in the Federal Agencies Digitization Guidelines Initiative's AV Working group [6], and monitoring the decisions of other large archives and library repositories. We learned that while many espouse best recommendations for preservation-worthy file formats, the outputs in real world production environments are often determined by practical decisions about conversion equipment and local technical capacity.

We currently exert little control over creating our current "master" digital format, MPEG-2. The MPEGs are acquired by using low cost, commercial desktop software to rip the access DVDs made from either a BetacamSP or polyester film copy. All downstream derivatives are created from this file. We have limited software and hardware for either in-house production or rendering of vendor-produced content encoded in many of the recommended preservation formats. Our current use case for MPEG-2 masters more closely aligns with NARA's Video Median Capture – SD profile, that of a reproduction master copy but not suitable as a substitute for the original. [7] This also parallels our existing preservation reformatting approach, where we equate a relatively high-end digital mezzanine file with the analog BetacamSP copy master.

Transcriptions and Captioning

Creating accurate transcripts and captions to satisfy Section 508 accessibility requirements proved to be the most time-consuming task in our workflow. Current resources limits our production capacity to about 50 titles per year. We initially hoped to get close to 100% transcription accuracy using modern speech recognition technology. However, experiments with automated audio extraction software such as Adobe Soundbooth only yielded accuracy rates with a mean of 51%. This is due to the poor audio fidelity of most analog film and video, background music, or ambient noise. Accuracy improved to between 70-90% with high-fidelity contemporary analog video or 2K/4K digital audio as represented by NIH's modern digital videocast productions. However, editing even these automatically generated high accuracy files to achieve the most accurate results takes more effort than a completely manual transcription process. Time-motion experiments with an application such as Dragon Naturally Speaking did not significantly increase transcript production Therefore, we currently use a combination of vendor-supplied transcripts when funding allows and transcripts created manually in-house. Manual transcription also allows us to have stricter control over the transcript formatting for import into captioning software, and can provide more subject-based knowledge for technical medical terminology, hard to hear passages, and foreign and other language issues.

These processes are time consuming and present a significant hurdle for any online film digitization project, however the access benefits, above and apart from satisfying the

DOI: 10.2352/issn.2168-3204.2016.1.0.117

Section 508 mandate, clearly outweigh the production cost factors. In the end it is more cost effective to outsource transcription and captioning production when budgets allow. Our experiences produced a wide range of time/motion metrics. The fastest rate any of our staff achieved was a 5:1 ratio per hour of runtime: 3 hours to transcribe 1 hour of runtime; 1 hour QA review (error-free transcript, no editing required); 1 hour captioning. This metric drops significantly for non-expert staff such as students and technicians -- closer to an 8:1 ratio per hour runtime.

Other variables, such as film content type, also impacted transcription time. Multiple-hour interview films that naturally have more speech take significantly longer to transcribe than a narrated training film. Another significant factor is the ability of staff to perform this work for a long stretch of time -- our experience is that production wanes after about an hour.

Broad Access and Derivative Creation

When we initiated our motion picture digitization project in 2010, it was important to satisfy a broad range of access use cases. Our initial content model, above, reflects the variety of access formats we selected.

However, we were probably over-ambitious in attempting to satisfy too many non-native use cases and file download audiences by supplying such a large number of other file types, especially considering that in 2010, NLM did not have one single application that could derive all of these formats.

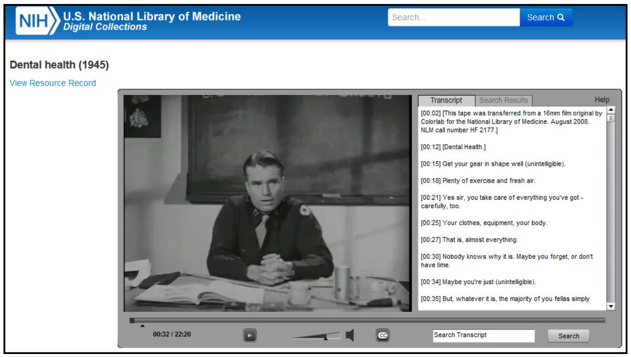
We anticipated that Mac and Windows users would want different codecs to download for local re-use, hence the highly encoded Quicktime (MOV), Windows Media (WMV), and H.264 (MP4) offerings. Those with non-flash devices such as tablets and smart phones would likely prefer file sizes optimized for those devices, hence the IPhone derivative. Magpie [8], our original captioning software, also required either a MOV or Flash file, and the video derivative tools at our disposal (Quickmedia Converter and Quicktime Pro) produced many of these files simultaneously, so it seemed to be more effort to discard the extra files than keep them. Magpie exported both DFXP and SMIL caption files, so it was little effort to also provide the SMIL files for those non-native MOV users.

Several other derivatives were created to assist users in navigating our primary user interface, and to aid potential future re-work. Still images in three different resolutions were created to represent the film in search results and detailed resource views. A thirty second curated clip was created in order to provide a quick preview of the film to the user. We also elected to keep the MAGPIE project file for each film in our repository, to aid in potential future re-processing of the video.

Video Search

NLM wanted to leverage highly accurate transcripts to provide a rich search, navigation and viewing experience for its digitized motion pictures. Rich search involved the ability to search full-text transcripts, and use these transcripts to discover films within the repository, and to discover the time location of the spoken text within the film. Navigation should include the ability to easily seek within the film to the identified search result locations. Standard video playback, navigational, volume and captioning controls should also be included.

To meet these needs, NLM developed NLM Video Search (NLMVS), an open source video player that provides rich search, access and navigational capabilities. We anticipated that most users would view our digitized motion pictures natively through our NLMVS player. This application was developed in-house at NLM using Adobe Flash and Actionscript, The player provides a video playback window alongside a pane for displaying the full transcript with associated time codes.



Dental health (1945)

[00:02] [This tape was transferred from a 16mm film original by Colorlab for the National Library of Medicine. August 2008. NLM call number HF 2177.]

[00:12] [Dental Health.]

[00:15] Get your gear in shape well (unintelligible).

[00:18] Plenty of exercise and fresh air.

[00:21] Yes sir, you take care of everything you've got - carefully, too.

[00:25] Your clothes, equipment, your body.

[00:27] That is, almost everything.

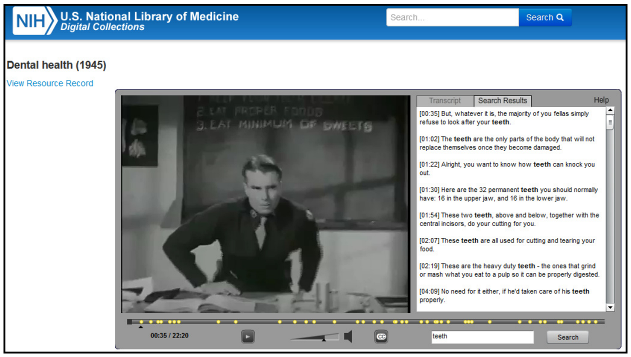
[00:30] Nobody knows why it is. Maybe you forget, or don't have time.

[00:34] Maybe you're just (unintelligible).

[00:35] But, whatever it is, the majority of you fellas simply

Figure 1. *Sample resource viewed with NLM Video Search player application*

A search function allows a user to keyword search the DFXP caption file. Search results with relevant time stamps are indicated by a yellow dot on a scroll bar below the playback window. Clicking any dot will take the user to that word's place within the video playback; search result snippets are also displayed in the transcript pane alongside the original full transcript.



Dental health (1945)

[00:35] But, whatever it is, the majority of you fellas simply refuse to look after your **teeth**.

[01:02] The **teeth** are the only parts of the body that will not replace themselves once they become damaged.

[01:22] Alright, you want to know how **teeth** can knock you out.

[01:30] Here are the 32 permanent **teeth** you should normally have: 16 in the upper jaw, and 16 in the lower jaw.

[01:54] These two **teeth**, above and below, together with the central incisors, do your cutting for you.

[02:07] These **teeth** are all used for cutting and tearing your food.

[02:19] These are the heavy duty **teeth** - the ones that grind or mash what you eat to a pulp so it can be properly digested.

[04:09] No need for it either, if he'd taken care of his **teeth** properly.

Figure 2. *NLM Video Search player showing search result with search result snippets, bolded search term, and time stamp dots on scroll bar.*

As a baseline, we wanted NLMVS to provide a high-quality video and audio experience but also satisfy the lower bandwidth capacity of a home desktop user accessing the internet with a broadband connection and without the benefit a streaming media server on our side to mediate data download time. Progressive download of a 480x320 pixel MP4 encoded at 300 Kbs and 22.1 bit AAC audio is produced as a compromise; this is the "Video Player" derivative noted in the content model above.

The NLMVS received an award from HHS Secretary Kathleen Sebelius as one of six winners [9] in an HHSinnovates contest, and is listed on the HHS IDEA LAB site [10].

Current Posture and Future Directions

We have since significantly altered our workflows and simplified our production processes. After a review of usage statistics and a move to a new Blacklight discovery interface in 2013, we adjusted our data model and removed several derivative files that we felt no longer provided significant benefits compared

DOI: 10.2352/issn.2168-3204.2016.1.0.117

to other available alternatives: Quicktime .mov (and the paired SMIL files), Windows Media .wmv, iPhone .mp4, and Clip .mp4. With a Java 7 upgrade the Magpie captioning software became incompatible with our computing environment and we changed to MovieCaptioner, a low-cost commercial application that offered most of Magpie's same functionality. We also exchanged the proprietary Magpie project file for SRT as a preservation master caption file. The library also purchased a networked Sorenson Squeeze that enabled us to create all the video derivatives at once using the MPEG-2 master. The revised content model includes the following items:

Video files, including:

- **Master MPEG-2**, native size (.mpg)
- **H.264** 640x480 derivative (.m4v)
- **Video Player**, 480x360, H.264 derivative (.m4v)

Caption files, including:

- **SRT** caption master file (.txt)
- W3C **DFXP** caption file (.xml)
- Plain **text** transcript file (.txt)

Metadata files, including:

- **MARC XML** (derived from ILS; definitive descriptive metadata store)
- **DMDINDEX** (local XML descriptive metadata transformed from MARC XML, for repository SOLR index)
- **Dublin Core**

Still shots, including:

- 640x480 **large** poster image (.jpg)
- 320x240 **medium** poster image (.jpg)
- 160x120 small poster image (.jpeg)

Moreover, shell scripts are invoked as pre-ingest processing steps that now automatically transform the DFXP caption file into the plain text transcript, resize the large poster image into the two smaller size images, and normalizes file names to conform with data model expectations. This reduces the number of files HAV staff produce and edit from 14 to 6. Finally, previous manual quality control steps are now mostly automated. A video QC module was created within our in-house Digital Projects Quality Control System (DPQAS) used for book scanning projects. MediaInfo [11] analyzes the pre-ingest SIP, and using the XML view, a shell script compares pre-defined values against the MediaInfo outputs and reports success or failure. Other scripts ensure expected file naming conventions are followed, and report any malformed XML characters in the DFXP caption file. These automation and QC steps significantly reduced ingest errors and re-processing efforts.

Most significant is the move from analog to digital as a preservation activity. Our access projects have given us more confidence to define requirements and select digital formats to replace our analog BetacamSP copying practice. Work will soon begin on a 100 title pilot to digitize U-matic video within NLM's General Collection holdings. U-matic tape represents our highest priority in terms of format fragility and hardware obsolescence. The long-term goal is to maintain our current production rate of 700-800 titles per year for both GC and HAV content. We have partnered with AVPreserve for technical and business advice, and proposed a tiered approach for film and video formats that would satisfy preservation requirements that also align with our enterprise storage, hardware, and software environments.

For master preservation formats we have decided to digitize video stock in a 8 bit Standard Definition 4:2:2 stream to a lossless, compressed FFV1 codec in a Matroska wrapper; film stock with little intrinsic significance as cultural artifacts will be digitized in a 10 bit High Definition 4:4:4 stream to an uncompressed AVI codec in a Quicktime wrapper. NLM has few culturally significant film artifacts and we expect these would be preserved film to film. The work would be largely outsourced--digitization for both preservation and derivative access files, and transcription/captioning. A similar DPQAS video module will be used for quality control, validation, and acceptance testing. All objects will be ingested into the NLM Digital Collections repository for long-term preservation management. Access to copyrighted or restricted materials will be gated in some fashion, by applying security controls and/or as a dark archive.

Conclusion

This article concisely describes our initial and current policies, workflows, and software used to develop online access to digitized motion picture content. Online film service is still a complex prospect. Other institutions have produced greater numbers of online film titles, but NLM has decided to provide value-added products for its users. High standards require deep commitment to programs that we hope best serve our public mission.

Acknowledgements

The authors would like to thank former CSC contract staffer Ed Luczak for his significant contributions in developing video content models and transcription workflows, in researching, optimizing and generating video derivatives, and in researching speech recognition technologies and approaches.

This work was supported by the Intramural Research Program of the National Institutes of Health, National Library of Medicine.

References

[1] https://collections.nlm.nih.gov/, NLM Digital Collections, accessed Dec. 6, 2015.
[2] http://resource.nlm.nih.gov/9300763A; http://resource.nlm.nih.gov/101306230.
[3] https://www.w3.org/TR/ttaf1-dfxp/
[4] http://www.archives.gov/preservation/products/definitions/reformatting.html.
[5] http://www.digitalpreservation.gov/formats/index.html.
[6] http://www.digitizationguidelines.gov/audio-visual/
[7] http://www.archives.gov/preservation/products/products/vid-r1.html Video Median Capture – SD, accessed June 18, 2012.
[8] http://ncam.wgbh.org/invent_build/web_multimedia/tools-guidelines/magpie. NLM does not endorse any of the products or companies discussed in this paper.
[9] https://www.nlm.nih.gov/news/nlm_video_search.html
[10] http://www.hhs.gov/idealab/projects-item/getting-more-out-of-video-nlm-video-search/
[11] https://mediaarea.net/en/MediaInfo.

DOI: 10.2352/issn.2168-3204.2016.1.0.117

Author Biography

John Rees is Archivist and Digital Resources Manager, Archives and Modern Manuscripts Program. He also serves on NLM's Digital Collections repository development team, shepherding film digitization projects and chairing the Preservation Working Group. He received his MLIS from the University of Texas-Austin, a MA in Southern Studies from Ole Miss, and his BA from Mary Washington College.

John Doyle is a Senior Systems Librarian in NLM's Library Technology Services Section. He is NLM's Digital Collections repository project manager. He received his MSI from the University of Michigan and BA from the College of William and Mary.

Doron Shalvi, CSCRA contractor, is a software developer and the principal architect for NLM's Digital Collections repository. He received his MS and BA in Electrical Engineering from the University of Maryland.