AMIA Tech Review
From the Association of Moving Image Archivists

Mav. 2012 | Issue 4

Video Preservation for the Millennia

By Linda Tadic, Executive Director, Audiovisual Archive Network

In the area of Dunhuang along the Silk Road in China's Gobi Desert are over 700 caves carved out of mountains by Buddhist monks. These caves, created between the 4th and 14th centuries AD as acts of devotion, are covered with incredible mural paintings depicting Buddhist sutras, details of daily life such as farming, dances, and ceremonies, and iconographic images from many of the world's religions: pagan Chinese, Buddhist, Hindu, and Christian. Some cave walls even include graffiti from Russian soldiers imprisoned in the caves in the 1920s.

The caves are managed and conserved by the Dunhuang Academy, which since 2000 has been creating high resolution digital photography and video documentation of the cave murals. The Academy is in an unending struggle to conserve the caves, which are a UNESCO World Heritage Site. Several caves collapsed over the centuries, succumbing to the sand, wind, and elements. The Academy knows that in 1,000 years, the digital photography could possibly be the only remnant of some caves and are developing a digital repository for the images, video, and data created. The digital files must be preserved as carefully as the caves themselves. I have been fortunate to work with the Academy in developing the requirements for their repository, and have been struck by how their experience parallels those of archives with videotape collections. While video has not existed for 1,500 years, the problems in conserving documentation of the caves and content captured on video are the same.

The Digitization Imperative

By the time today's moving image preservation students reach their 25-year career mark (if not sooner), videotape will have ceased to exist. Not only will the manufacture of videotape have ended as the world completes its shift to file-based production, but most of the legacy videotape held in archives will no longer have a retrievable signal to transfer. Like a Dunhuang cave that could be subsumed by the elements, the content on those videos fortunate enough to have been digitized will survive as digital files; in essence, the video content is "re-born" as a digital file, and it is the digital file which must be

preserved for the future. With video, archivists have had to shift their focus from preserving the medium to preserving the content on it.

Videotape was never intended as a long-term storage medium. Its inherently short life expectancy (LE) required archives in the pre-digital era to periodically migrate from one video format to another in an effort to prolong the content's existence. Compounding the repetitive migration problem was video format obsolescence, which every archive with video has experienced. Twenty years ago, transferring to Umatic was the standard operating procedure. Umatic, now an obsolete format, was followed by BetaSP as the standard target analog format. BetaSP is now obsolete. Archives began transferring to Digital Betacam, but while the media holds a digital signal, the content is still on videotape with its inherent deterioration issues; the file must be extracted from the tape and treated as a digital file. In addition, archives are beginning to consider DigiBeta as the format next in line to become obsolete. Its high definition (HD) sibling, HDCAM, is commonly used by broadcasters and studios for HD content, but again this is a videotape-based format. The HDCAM tape shortage caused by the March 2011 earthquake and tsunami in Japan acted as a "wake up call" for content creators to stop delaying the inevitable and make the move to medialess production.

Target preservation digital format?

With video's format obsolescence, short life expectancy issues, and the impending death of video as a viable medium, content creators and caretakers of all types are recognizing that they must adapt their production and archiving workflow from analog to digital. Content is being created digitally ("born digital"), and analog videos are being digitized as funds allow.

The process of transferring legacy obsolete video formats to soon-to-be-obsolete video formats is in danger of being replicated in the digital realm. Transferring to digital codecs or formats that can possibly become obsolete in the future is not a good model to follow.

One of the most contentious topics in our field is what should be the standard target preservation digital format when transferring analog tape to digital files. Some believe that there cannot be one standard appropriate for all analog source video formats and for all archives; instead, we should focus on using an open format for easier future migration and consider the organization's infrastructure and ability to preserve the digital files.

In selecting a target format, one must weigh file format sustainability factors such as those outlined by the Library of Congress.[1] The basic rule of thumb is that the preservation format should be:

- An open standard (not proprietary; this includes file wrappers as well)
- Well-supported (strong hardware and software support)
- Well-documented (required so validation and other tools can be created to check the file)

Ideally, the file should be as uncompressed as the archive can manage. Using an open format with as little compression as possible will help an archive migrate the files forward in the future. Low or no compression is also more forgiving of bit loss, whereas bit loss in a highly compressed format can result in lost information or even a corrupt and unplayable file depending on where the loss occurred in the file.

The **open standard** requirement means there are few file formats or codecs from which to choose for analog-to-digital conversions. The most common is uncompressed (YUV 10-bit), followed by JPEG2000. While DV25 is compressed 5:1, it and DVCPRO50 (3:1 compression) are sometimes used when transcoding VHS tapes or by archives with such large video collections that their infrastructure could not support encoding everything as uncompressed.

Sustainability factors for wrappers such as MXF (an open format unless proprietary information is added to the header) and the two proprietary wrappers QuickTime and AVI must be weighed just as much as a codec or format. The Federal Agencies Digitization Guidelines Initiative **(**FADGI**)** is drafting an open MXF Application Specification for Archiving and Preservation (AS-AP)[2], and the result of the group's work is eagerly awaited by the archival community.

Long-term storage

In choosing a preservation digital format, the organization must consider its infrastructure. Can the organization support the digital storage, staff time in migrating and checking the files every set number of years, and upgrading hardware roughly every five years? We hear that "storage is cheap," but storage itself is only one part of the ongoing digital preservation costs. Before selecting a target preservation format, an archive must estimate the amount of storage required for at least five years' growth, as well

[1] http://www.digitalpreservation.gov/formats/sustain/sustain.shtml
[2] http://www.digitizationguidelines.gov/guidelines/MXF_app_spec.html

as human labor and infrastructure costs. Infrastructure can include hardware, software, electricity, air conditioning, physical space, and backup generators.

There is no "store and ignore" media for digital files. Digital preservation requires constant managed actions to migrate forward both the storage media and the file format itself should it become obsolete. This work must be done regardless of the storage medium used: external hard drives (HDD), RAID servers, a Storage Area Network (SAN), or digital tape such as LTO. Larger organizations often use a mixture of storage strategies, for example using a SAN with automated LTO backup. Smaller organizations tend to use HDDs or standalone RAID servers, but it is becoming more common for them to also make LTO copies for backup using single-slot LTO drives. None of these are perfect solutions, so an organization must research and understand the pros and cons and the work involved with each storage solution.

A SAN with automatic LTO backup can involve higher costs upfront for hardware and infrastructure (electricity and AC) with lower staff costs for ongoing maintenance. A smaller operation utilizing HDDs or a RAID attached to a single-slot LTO drive will have lower initial hardware and infrastructure costs, but be higher on labor since the backup process might not be easily automated. Additional considerations in using LTO as a back-up medium is that the tape is manufactured to be two generations backwards "read" compatible, and one generation backwards "write" compatible. This means that LTO3 tapes may be read on today's LTO5 decks (but not written to), but LTO2 tapes cannot be read or written to; LTO2 is an obsolete digital tape format. Thus, an archive using LTO is obligating itself to upgrade hardware every two generations (approximately every 5 years). There is also the issue of back-up software, which writes and catalogs the files to tape so the files can be retrieved. If an archive sends content to a vendor without specifying which backup software they use, it is possible that the vendor could return the LTO tape with their transferred content using a backup/cataloging software that is not readable by the organization. LTFS (Linear Tape File System) is a setting now available on some LTO5 decks that makes LTO tape behave like simple file storage on a hard drive; this is a promising development in removable file storage.

External hard drives (HDDs) should also be refreshed every 3-5 years. HDDs are an inexpensive storage medium for smaller archives, but are notorious for failure between 3-5 years of use. Just as files on LTO

tape must be migrated every two generations, files on HDDs must also be migrated to new devices every 3-5 years.[3]

Content protection

Regardless of the long-term storage implemented, a digital preservation strategy or plan must be enacted for content protection. At its most basic, the preservation plan should include capturing data on file creation at the very beginning of its lifecycle, file fixity checks (checksum), redundancy and geographic dispersal, and scheduled storage migration (with file format migration as formats become endangered or obsolete).

The **file creation information** is called technical metadata; this can be used to preserve the file in the future. Everything about creating the file must be captured: the hardware and software used if it was an analog-to-digital transfer, or the camera/device used if it is a born digital file; the technical characteristics of the file (codec, format, version, size, etc.), and the environment in which the file can be rendered/played (e.g., which browser version, which software program and version, etc.).

The **file fixity check** is performed through a checksum. A checksum is created when a file is first born, and is run to check for bit loss every time a file is transmitted from one storage device to another. The checksum creates an alphanumeric string that is unique to that file; if the checksum does not match after transmission, then the file has been somehow corrupted. The most common checksum algorithms are MD5, SHA-1, and SHA-2. MD5 is the most common checksum but is now considered the least secure, so many organizations are moving to SHA-1 and SHA-2.

The most important act in digital preservation is **geographic dispersal** of multiple copies **(redundancy**). A file can be corrupt on one LTO tape or HDD, but is fine on another. Storage devices are not infallible: a tape can be creased or warped, the drive on a HDD can be damaged by fine particulate matter or from vibrations. If an organization can afford it, three copies are recommended but at a minimum there should be two. A redundant copy does not mean one copy is in the basement and the other on the 3rd floor; it means the copies are placed far apart from each other. If a fire destroys your building, your

[3] See Google's study on hard drive failures. Pinheiro, Eduardo et al. "Failure Trends in a Large Disk Drive Population." *Proceedings of the 5th USENIX Conference on File and Storage Technologies (FAST'07),* February 2007.

content would be safe at other facilities, say in Colorado and Pennsylvania. Archives could have cooperative arrangements where one archive stores backup copies for the other and vice versa; this storage would have to be secure so there is no risk that the copies could be stolen or damaged.

Migration to new storage media should be scheduled every 3-5 years, depending on an organization's choice of storage medium. The larger the file, the longer it takes to retrieve the file from storage, check it, and copy it to a new storage device. If proprietary file formats/codecs are used rather than open ones, the obsolescence arc of that format must be watched so the archive can migrate the file format forward as necessary. An archive's database should track the codec and wrapper of every file so reports on endangered or obsolete formats can easily be run. All of these migration actions must be tracked through metadata.

Metadata

Human labor can be the most expensive component of any digital preservation strategy, and part of the human labor is metadata creation. The previous section described digital preservation actions, and metadata must be created to track every one of those actions. Much of technical metadata can be automatically extracted from a file, but a human must input the majority of descriptive information, details on how a file was created and migrated, and intellectual property rights.

The core components of a metadata record are: descriptive (what the content is about and who was responsible for creating it), rights (who owns the rights to the overall work, and what are the underlying third-party rights), technical, and preservation. There are data structure standards for descriptive, technical, and preservation metadata from which an organization can choose relevant fields. No longer is there one over-arching data structure such as MARC; today, an organization chooses the fields from various standards most relevant to their collection and creates a data dictionary incorporating the fields. A data dictionary includes a map between the various standards for interoperability with other collections and future data migrations.

Data structure standards useful for video can range from basic (Dublin Core), broadcasting-oriented (PBCore and EBU Core), film archives-specific (CEN 15744 and 15907[4]), film delivery-oriented (SMPTE

[4] "CEN" stands for Comité Européean de Normalisation

DMS-1), and of course there is still MARC. Technical metadata standards include SMPTE's RP-210, and PBCore and EBUCore's technical metadata set. Digital file preservation metadata is only represented by PREMIS; this is a format-agnostic standard. An organization should try to create as granular[5] a data structure as possible for easier future data migration; it is inevitable that data will be migrated to new systems several times over decades.

Controlled vocabularies (e.g., set terms for subjects, names, and places) should be used to ensure the data is consistently described and content can be easily retrieved by users. Controlled vocabularies can take the form of simple "pick lists," thesauri, hierarchical taxonomies, and synonym rings.[6] If an organization does not use a standard controlled vocabulary, it should create an internal one. The key concept is to be consistent.

In the digital asset lifecycle workflow, metadata creation tends to be performed by more than one person. One staff could add basic descriptive metadata at the beginning of the asset's lifecycle, another will add technical metadata, another add rights information, and the library/archive will add preservation metadata and perhaps controlled vocabularies. Utilizing several staff to create metadata can alleviate the workload, but smaller organizations often have just one cataloger/metadata librarian.

Time

Archives know the clock is ticking against them and much video content will inevitably be lost. They make hard decisions on what they can afford to transfer in terms of file storage and ongoing preservation actions. To save at least some representation of the content, some archives feel that making highly compressed copies are better than not transferring the video at all. Video preservation is a complex process, and to ensure that video-based content will last as long as the Dunhuang caves,

[5] "Granular" in a data context means there are discrete fields holding specific information, rather than one field containing several data elements. For example, using separate fields for "format," "running time," and "manufacturer stock" instead of compiling all three elements into one "note" field will facilitate mapping the elements to a new database.

[6] For guidance on creating controlled vocabularies, see: National Information Standards Organization (NISO). *Guidelines for the Construction, Format, and Management of Monolingual Controlled Vocabulary*. ANSI/NISO Z39.19-2005. http://www.niso.org/kst/reports/standards?step=2&gid=None&project_key%3Austring%3Aiso-8859-1=7cc9b583cb5a62e8c15d3099e0bb46bbae9cf38a

video archivists must become digital preservationists. We are experiencing not just a shift in how to preserve video content, but also a shift in our skills as preservationists and archivists.

Glossary

MXF (Material E**x**change **F**ormat). This is a SMPTE standard (*SMPTE ST 377-1:2011, Material Exchange Format (MXF) -- File Format Specification*). MXF is a container or wrapper format that can hold several bitstreams such as video, audio, and XML. Uncompressed YUV and JPEG2000 files are supported in MXF. MXF is the container for Digital Cinema Packages, and for files created on Sony XDCAM and Panasonic P2 cameras. See: http://www.digitalpreservation.gov/formats/fdd/fdd000013.shtml

QuickTime. Apple's proprietary container or wrapper format, also known by its file extension MOV. This is the native wrapper for Final Cut Pro editing software. Uncompressed YUV files are supported in QuickTime, but not JPEG2000. See: http://www.digitalpreservation.gov/formats/fdd/fdd000052.shtml

AVI (Audio **V**ideo **I**nterleaved)**.** Microsoft's proprietary container or wrapper format. Uncompressed YUV files are supported in QuickTime, but not JPEG2000.
http://www.digitalpreservation.gov/formats/fdd/fdd000059.shtml

LTO (**L**inear **T**ape-**O**pen). A form of digital tape that stores data. While it is magnetic media, the formulation of data tape differs from videotape. There are a few competitors in the data tape market such as DLT and AIT, but in Q1 2011 LTO had 87% market share so only this product is mentioned in this article. LTO5 is the most recent generation of LTO, and can hold 3 TB uncompressed data or 1.5 TB compressed. See the LTO roadmap: http://www.ultrium.com/technology/roadmap.html

RAID (Redundant **A**rray of **I**ndependent **D**isks). A server configuration for how the server or external hard drive protects files. Levels are referred to by numbers. RAID0 = no redundancy; the server or HDD is pure storage with no internal redundancy. RAID1 = files are "mirrored" (e.g., copied) from one internal drive to a second. RAID2 through RAID6 use striping and parity; for a full explanation see the Wikipedia entry here: http://en.wikipedia.org/wiki/RAID

Sources for metadata standards referenced

MARC21 http://www.loc.gov/marc/

Dublin Core http://dublincore.org/

PBCore http://www.pbcore.org/

EBU Core http://tech.ebu.ch/docs/tech/tech3293v1_3.pdf

CEN 15907: Film identification — Enhancing interoperability of metadata — Element sets and structures; prEN 15907:2009 http://filmstandards.org/fsc/index.php/EN_15907

CEN 15744: Film identification — Minimum set of metadata for cinematographic works http://filmstandards.org/fsc/index.php/EN_15744

The SMPTE standards can be acquired from the SMPTE site: http://store.smpte.org/