

Archiving Strategies for Long-Term Video Preservation

Steven D. Rubin
January 2009

Digital video has become, within the last several years, a mainstream communications method. Every day, thousands of corporations, groups and individuals now create video that is viewed online by millions. Television has completed a transition to digital transmission, and digital cinema will soon replace film in your local theater. A political campaign is not complete without online access to commercials, speeches and events.

The ubiquity of digital video is evident with any look at the numbers. In September 2008, for example, YouTube alone counted 81 million unique visitors who retrieved more than 5 billion video streams.¹

While most digital video being produced today does not require long-term archiving, certain video files do need to be preserved for the long term. Some video files need to be maintained for legal reasons, such as corporate investor relations webcasts. Others, such as digital TV programs, will be archived for commercial and historical reasons.

The sheer amount of storage itself is impressive. An hour-long HDTV program, for example, typically consumes about 100 GB of storage space.² When one considers the number of programs released each day, accumulated over years, the number of programs and hours that must be archived is considerable.

These forms create a challenge for the archivist who must deal with a number of considerations at the onset of the archive process that have long-term, downstream consequences.

Video Format Considerations

A media format is a package containing several different parts, including the video compression scheme, or codec, and audio codec, and a file format. It is very common for video to be created in one format (such as MPEG-4), and “transcoded” so that an end-user views it in a format that is more readily streamed to desktop computers (such as Adobe Flash).

Fortunately, many of the most common video formats rely on the same set of industry-standard parts. These are increasingly clustered around the use of a single video compression standard, H.264. (A compression standard is simply a methodology for reducing size of a video file so it

¹ CNET news, Digital Media, October 31, 2008, http://news.cnet.com/8301-1023_3-10080076-93.html

² Mary Ide, Dave MacCarn, Thom Shepard, and Leah Weisse
WGBH Educational Foundation, Understanding the Preservation Challenge of Digital Television,
Council on Library and Information Resources,
<http://www.clir.org/pubs/reports/pub106/television.html>

can be more easily delivered or stored). Common formats such as MPEG-4 Part 10, Flash and Quicktime all are based on the H.264 standard. (check into others such as WMV).

However, formats with the same standard will typically compress video at different rates, greatly affecting the resolution and overall quality of the file. For example, (bit rates for MPEG 2, MPEG 4, Flash).

Therefore several choices need to be made, with the media format itself as the primary consideration. If the media must be maintained for many years, the predicted longevity of the format will be a key factor. Generally, formats that are deploy industry standards with little or no proprietary technology would seem to have the best chance of surviving technical evolution. For long-term preservation, media should be stored at the highest level of quality and resolution. Lower-resolution derivatives can be created from the originals, but of course there is no way to derive high-resolution media from a low-resolution archive. The compression rate within the format should be evaluated and considered.

If the volume of digital media is composed of mixed formats, another consideration is whether to maintain each in its original, or convert all to a single format. Although it is appealing to the archivist to retain the an asset in the format it was created, this might not be the best method of preserving it for the long-term. If the quality of the original can be retained, it might be better to convert the entire volume into a single format that is considered to be the best bet for long-term survival.

In the end, there is no guarantee that any one format will exist into perpetuity, or that a format will be easily convertible a form that allows it to survive into the next evolution in technology. A decision about media formats is, unfortunately, speculative by nature, based on an informed prediction of the future direction of media and content management technology.

Repository Strategy

The repository of an archive system is the software-based container that holds the media files. Repositories tend to range from open source software, to proprietary software that is tied to a digital asset management system. A decision regarding a repository is often driven by the more immediate needs of the end-users. For example, an organization that has production-oriented needs for workflow or multi-layered security is more likely to seek a full-featured digital asset management system that supports such features, with less attention paid to the underlying repository.

A DAM owner will prioritize these functions according to their business needs. For example, a marketing organization involved in product launch activities will lean heavily on the workflow and versioning functions of the application, and might be less concerned with the long-term preservation of content as the product changes or sunsets. A non-profit organization seeking to preserve its history, on the other hand, will be very concerned about the longevity of assets, but workflow and versioning might be less important.

Long-term, continued access to assets becomes a more important factor as the business needs tend toward preservation, rather than production and distribution of content.

Most modern DAM systems are based on a tiered architecture that enables a distinct repository to be accessed by a number of creative and workflow applications. **Any repository strategy, therefore, must take into account the overall and short-term need**

for the production-oriented features of modern digital asset management systems, balanced against the long-term need for accessible archives.

There are several approaches which each have pose distinct advantages and challenges:

- **Open Source repositories** that are not dependent on any one company and are only loosely tied to technical standards. There are two dominant open-source repositories in use, Fedora and JSR-170. Fedora emerged from largely from the academic community. It was created by Cornell University as a project of the National Science Foundation and DARPA. It offers a multi-media repository in which metadata is tightly linked to the media file. As open source software, development is distributed, and coordinated through a central organization known as Fedora commons.

JSR-170 is a specification of the Java Community Project, largely used as a depository for web content management systems. In other words, it is an industry standard in which a software developer can comply with.

The problem posed by both of these open source standards has been that few firms develop applications or tool sets on these platforms that make the repository more useful in the short-term. In other words, although an open source might provide a good long-term solution, its often provides limited short-term usability.

- **Widely-accepted proprietary repositories.** Several software vendors offer data repositories that are in such widespread use that they can be considered a de facto standard. One such example is Microsoft's SharePoint, which has many thousands of users, along with a community of qualified developers. In order to extend the usability of SharePoint, a number of applications and extensions are now on the market.

The advantage of this approach, then, is that a number of applications might be available for the same platform. Some of these might be customized to meet somewhat granular requirements. Further, older applications can be more easily retired and replaced with no loss of data and no need for migration.

Despite the wide use of the platform, a proprietary repository does pose a potential continuity risk, since development and support can be withdrawn as the software vendor changes strategy. Take note, for example, of IBM Content Manager, which had nearly 9,000 corporate customers at its peak. After IBM's purchase of FileNet in 2007, development shifted away from CM to FileNet's platform, P8. IBM's lead business partner for digital asset management, Ancept, followed by re-writing its application to support P8, stranding older customers on the previous platform.

Proprietary depositories. A common approach to archiving is deployment of digital asset management software. DAM systems wrap sophisticated production, security, search and organizational functions around a central repository. In many cases, the repository is totally proprietary to the DAM application. While it may be only loosely linked in a layered software environment, and might be accessible through a web interface as well as the vendor's thick client, generally the repository is, realistically, tied to the application. There is little likelihood that independent developers will create applications and interfaces for a DAM vendor's platform. A more serious risk is that the platform quickly obsoletes in the case of the business

failure of the vendor, or if the vendor decides to abandon it as it seeks a more modern approach.

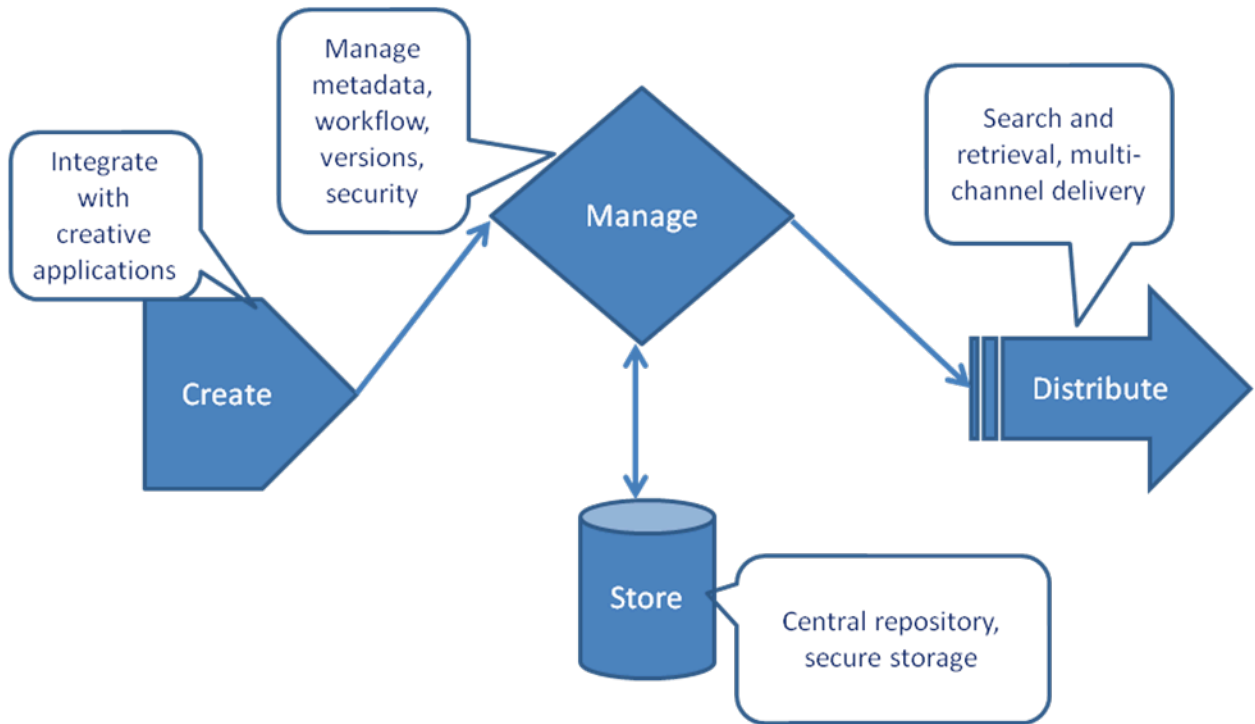


Figure 1 – Overall functions of a typical digital asset management solution. In general, the central functions – manage and store – represent standard archival functions. Integration with creative tools and distribute techniques, along with security, are value-add functions that distinguish software as a digital asset management application.

A minority of digital asset management vendors will now work with nearly any repository. MediaBeacon, for example, manages media files within the original file structure in the Macintosh or Windows environment and does not require an additional repository layer. Although some functionality might be lost compared to a more traditional DAM, the lack of a proprietary depository reduces the likely interval in which data must be migrated to a succeeding platform

Metadata and Taxonomy

There is considerable debate about how closely metadata should be linked with the video file format. The world of still images has moved decidedly toward imbedded metadata such as XMP, in which the metadata is part of the file format. Video archiving, for the most part, stores metadata as XML, loosely attached to the media file. This complicates data migration, but it allows for flexibility in the metadata schema as well as the types and numbers of fields.

Again, the best practice is to seek an industry standard, such as Dublin Core, rather than a proprietary metadata schema based on either vendor software.

Physical Storage Considerations

The half-century history of electronic data storage reveals that there is no such thing as a permanent data store. The first reason for this is that all storage media deteriorate over time. Magnetic pulses weaken, tapes disintegrate chemically, spinning disks crash. The second reason is that storage technology keeps changing. In fact, the rate that new storage technologies are introduced – and that older types are obsolete – has generally surpassed the rate of storage media deterioration.

The natural obsolescence of storage devices poses a potential danger for archives that have not been migrated. For example, although cinematic film can be preserved for decades and possibly centuries under the right conditions – cool temperature and very low humidity – projectors and telecines needed to view the film are not likely to be available.

Fortunately, storage also has become increasingly efficient and inexpensive. Therefore, a sensible strategy anticipates migration of the entire volume at somewhat narrow intervals (recent history shows about one decade as a reasonable life for a storage format). However, care must be taken during migration to preserve the file type in its original resolution, and to avoid compression techniques that do not allow a frame-accurate reproduction of video.

Conclusions:

- Video of all types is now a mainstream communications tool and being produced and consumed at increasing rates
- Much of this video will need to be preserved for years for legal reasons, or for decades for purposes of historic preservation
- Formats based on industry standards, such as MPEG-2 and MPEG-4, are safer choices for the long-term than proprietary formats.
- The organization must determine the importance of long-term archiving compared to short-term production and workflow considerations. Each will effect different repository strategies.
- Long-term archiving is not always compatible with proprietary repositories provided by digital asset management vendors
- Widely accepted and industry standard metadata schema, such as Dublin Core, are more likely to survive in the long-term
- All storage media types have a lifespan. An archive strategy should include plans for migration to new platforms over somewhat narrow intervals.

Steven D. Rubin, principal of SD Rubin Digital Media Strategies, is a consultant in digital media strategies, technologies and digital asset management. He can be reached at sdr@sdrubin.com, or 215-985-1777. www.sdrubin.com