Chapter 11
Digital Archiving in the Entertainment and Professional Media Market

Thomas M. Coughlin*

Abstract This chapter explores the current trends and drivers for digital archiving in the entertainment and professional media markets. Methodology and technologies utilized in digital archiving as well as future requirements will be explored. We will also examine the requirements for long-term content storage, look at various storage options and discuss how content administrators can create a reliable long-term archive.

Growth in the Entertainment and Media Market

To understand the demand for archiving in the Entertainment & Media (EM) market, it is important to begin with a discussion of the growth in content attained by all forms of EM companies from broadcast to post, animation to digital imaging and pre-press.

For companies participating in the entertainment and professional media markets their assets are primarily the rich content that they create and manage. A movie, a piece of music, a television program, a documentary, a video or audio file for web download and news or advertising pieces all generate revenue. They also contain large amounts of data and consume storage capacity. It is important to note that the raw data collected to make these assets is much greater than what will be used in an actual product.

For instance, a professional animal videographer told me that he shoots 10 h of high resolution video to finally get 5 min of selected and edited footage for a television documentary. Those 10 h of video were captured in 10–40 min periods over the course of many months while waiting in the open for hours at a time. There have been days when he did not get any images at all. Obviously, for content creation professionals there is a tremendous investment in the raw content.

* The material presented in this report was largely extracted from the *2007 Entertainment Content Creation and Digital Storage Report* or from the data set created in putting together that report. For more information or to order a copy of this report, please go to http://www.tomcoughlin.com/techpapers.htm.

D. Gerbarg (ed.), *Television Goes Digital*,

These pieces of content are also unique and can never be recreated exactly – time moves on. As a result, raw and edited content are of inestimable value in preserving a record of natural history, historical events, hits and fads of a particular age and generally serve as a cultural record of who we were and what mattered to us.

For the videographer to maintain this 10 h of raw content he would need 648 GB of storage at HD-cam resolution. Retaining raw content can get expensive as it is continuously acquired and maintained. Of course this problem is not unique to videographers. From broadcast to post to imaging, all media and entertainment artists have to deal with storage of growing collections of raw content.

The effect of this is that professionals are creating and storing more content than ever before. Figure 11.1 shows projections out to 2012 for the growth of annual storage capacity requirements for the creation of professional moving image content. In 2007, over 700 petabytes of storage hardware were required while in 2012 this is estimated to climb to about 2.4 exabytes.

One of the drivers of this increase in storage capacity is the move to higher resolution content. As the resolution of video content increases the digital storage required increases. Table 11.1 shows the bandwidth and 1-h storage capacity demands of typical professional media formats (including sound) and the primary applications that they are used in. The move from SD to HD has resulted in 6X higher data rates and storage capacities while 2–4 K has resulted in 4X higher data rates and storage capacities. Bandwidth increases result in greater production of storage content and thus more demand for storage capacity.

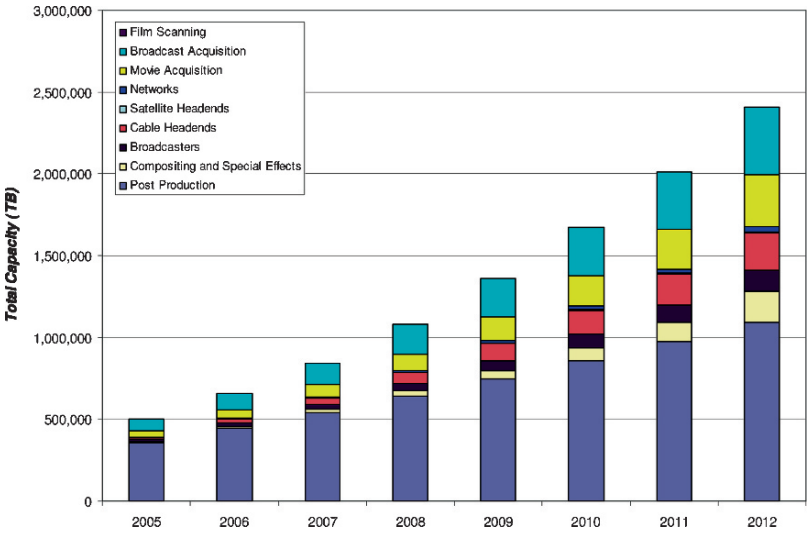

Fig. 11.1 Annual capacity projections for creation of professional moving image content. *Source:* T. M. Coughlin, *2007 Entertainment Content Creation Digital Storage Report,* Coughlin Associates, http://www.tomcoughlin.com/techpapers.htm

Table 11.1 Example raw data rates, professional media standards

Resolution	Application	MB/s	Storage (GB, 1 h)
Uncompressed audio (48 kHz/24-bit)	All	0.15	0.540
MiniDV/DVCAM/ DVCPRO (digital video)	Consumer, Corporate, Broadcast & News	3.6	12.96
IMX D10	Broadcast & news	6	21.6
HD Cam (compressed hi-def)	Broadcast, post production	18 (max.)	64.8 (max.)
SD Video (8-bit, uncompressed)	Broadcast, post-Production	20	72.0
Uncompressed 601 (SDI) (10-bit standard def.)	Broadcast, post-production	34	122.4
High Resolution HDTV (8-bit, 1080i)	Broadcast, post-production	120	432
Uncompressed 2 K (10 bit-log)	Film production, television	300	1,080
Uncompressed 4 K (10 bit-log)	Film production, some television	1,200	4,320

It should be realized that Fig. 11.1 gives an estimate for hardware storage assets required to make digital content in each year. The actual digital content generated using this hardware will tend to be even greater. Raw content stores alone can be considerable since a movie producer might edit hundreds of hours of content to come up with a 2 h final product. Then there is the unused content saved for outtakes or even sequels and the digital intermediaries and effects created during production. Producers, who make a practice of saving their entire source raw material can in some cases, end up with several petabytes of content.

When the digital content created and retained from prior years is added in, the cumulative digital storage required for a company to maintain its revenue generating content for the long term can become very great indeed. This is especially true as more ways to use and distribute captured content are identified (e.g., Internet and mobile phones) and the demand for richer and richer content increases.

Figure 11.2 shows estimates for storage capacity required to retain digital cumulative content created since 2004. These estimates assume that only a part of the total raw content is preserved. If the percentage of raw content retained over the long term is greater than assumed here the total cumulative storage capacity could increase by 50% or greater.

To these estimates must also be added the digital storage required for the retention of digitally converted, formerly analog content such as old movies and television programs as well as music and digital still images. Furthermore, extra copies of content such as are needed for disaster recovery will add to the total storage requirement. *As of 2007 the total unconverted analog video and movie content that could potentially be converted to digital form is on the order of 200 exabytes*.

The growth of converted content depends upon the assumptions made about the conversion rate and resolution requirements as well as the continued rate of

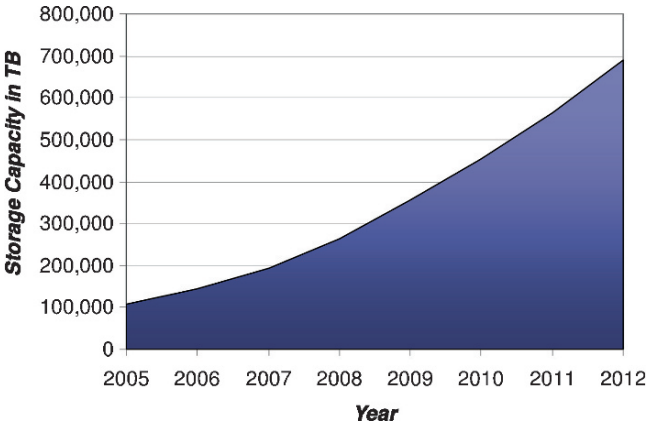

Fig. 11.2 Cumulative digital storage projection for professional video digital content preservation. *Source*: Produced from background data for T. M. Coughlin, *2007 Entertainment Content Creation Digital Storage Report*, Coughlin Associates, http://www.tomcoughlin.com/techpapers.htm

creation of analog content that may need to be converted. Together with copies made of content archives for disaster recovery the total digital storage required for the cumulative preservation of professional digital content by 2012 is estimated to be many exabytes.

As can be imagined, storing all of this content on disk is not feasible. For this reason, the entertainment and media industry is at the forefront of using archiving to cost-effectively retain and manage this content.

Archiving in the EM Market Defined

An *archive* is a copy of data that is being retained for very long periods of time, usually for years and in some cases centuries. Archives are used throughout the entertainment and media industry for storing content that is not being used in ongoing projects, but could be re-purposed or referenced in the future. An archive may be *active*, online, where it can be accessed relatively quickly or *cold*, offline, where it can be stored safely and economically; but it may take a considerable amount of time to mount the digital storage medium and read the archived data. The average time to access archived data is the archive latency.

Editing and some other content industry segments also keep *working archives* of content on storage networks during the course of their work. These working archives are raw content and edited content that are protected during active work on a project. They are often kept in storage area networks (SAN) or network attached storage (NAS) systems used in the working studio. After a project is completed, the content of a working archive may be retained in a long-term archive depending on

the value and time to create intermediaries, certain effects and other content used to develop the final cut.

Status of the Professional EM Archiving Market

Archiving within the EM market is currently driven by two factors: The need to cost-effectively retain content for re-use and the need to convert historical analog content to digital form to prevent degradation of content. Many sorts of facilities keep content for varying periods of time. Keeping completed content in long-term archives is common practice among content owners, including movie and television producers. Raw content retention is not so complete and varies depending upon the policies and budgets of the facilities. Post, special effects and computer generation houses may keep some of their unique content for extended periods. As the cost of archiving declines more content will be retained by all of these facilities.

For preservation of new content, deciding what to archive can be difficult. The retention of digital entertainment and professional media involves costs in real estate, operation, management software and hardware. In practice data managers handling large and even small volumes of raw and edited content must make choices in what they will preserve for the long term because they often do not have the resources or budgets to save everything.

Preservation of source media containing the original content is almost always a stringent requirement since it is difficult, if not impossible to recreate. However, this alone is often insufficient, as raw source material does not capture any edits or metadata generated during the processing of the raw content to create a finished product. As a result, more content must be archived than the source material.

To this end, many movie studios and editing facilities will save a set of master movie films (with color separation preserved) as well as managed digital tapes of the content in a cold archive. In addition, many of these facilities will keep copies of content on-line in an active archive on a disk array or tape library for some period of time.

Many television networks, including major news networks, retain their library of analog video tape with a digitized index and metadata database of their analog tapes. They convert the analog content to a digital form and store it on a combination of disk arrays and tape libraries as they are accessed and used for historical content for current projects. Networks tend to keep cold as well as active archives so that they can improve their odds of rapid content retrieval and successful long-term content retention.

Smaller organizations, including smaller post-production shops and other facilities, may not have the resources to set up a complex and expensive archiving system. For these users it is advantageous to retain the source media and players to play them back later. Careful logging and management of these physical media combined with retention of the most valuable content in disk arrays or tape libraries may be a more affordable option.

While archiving will be driven in the long term by new content creation, there is also a sharp increase in digital storage capacity used for digital conversion and preservation of analog historical content. Costs for digital conversion are being reduced with the development of service providers who can do the conversion of bulk material at an attractive price and with the decline in the overall cost of storing digital archives.. *We project that several exabytes of digital storage will be used for digital archiving and content conversion and preservation by 2012!*

Long-term retention of digital files is not merely a need of the entertainment and professional media industry. Regulatory and other pressures are driving many industries and companies to turn to long-term data retention as well. The Storage Networking Industry Association (SNIA) started a 100 year archive project in 2004 and in August 2007 it released a survey of business professionals speaking on their long-term archiving needs. In that survey it was found that 80% of respondents have information that they must keep for over 50 years, and 68% of respondents said they must keep this data for more than 100 years. Thus there is a strong incentive to develop hardware and management software to support long-term retention requirements for several industries.

Solutions and Products for the EM Archiving Market

Solutions and products for archiving vary depending on the overall size and needs of an organization.

For smaller organizations, archiving may be a simple process in which final cuts from working storage is written out in digital format to media and placed in a cold, offline archive along with source tapes. Products used for data movement are typically asset managers but may also include backup applications. For digital content written out from working storage, archive media uses “commoditized” IT storage like digital tape, optical storage and even disk drives and disk drive cartridges. These commoditized products could include LTO and DLT tapes, DVD or blue laser DVD write one or rewritable media as well as Iomega REV drives or disk drives in removable modules.

Factors in the choice of cold archive storage formats include not only cost, $/GB, but the expected life of the archive media. Media life depends upon the quality of that media as well as the environment that it is stored in. It is likely that inexpensive off-the-shelf optical disks, for instance, may not provide long-term data preservation (many tens of years). However, disk while being an easy to use format, is not suitable for cold archives because of its life span. For this reason, most cold archives use tape (e.g., LTO) due to its portability, quality and long shelf life – upwards of 30 years. In any event, media should be stored in a climate controlled vault for high quality preservation.

Larger organizations will also use cold archives but in conjunction with some form of active archiving. In real world deployments, active archiving tends to include

a hierarchical system consisting of multiple storage tiers. A disk storage array may be used on the front end to provide faster content transfers between artists during ongoing and recently completed projects. Less frequently used files will reside on a local tape or even optical disk library system for near line access.

Disk arrays used for active archiving typically use SATA hard drives since these provide higher storage capacity for a given price and provide good reliability under less frequently accessed conditions expected in an archive application. There are even SATA storage arrays called MAIDs (*massive array of inactive disks*) that keep most of the SATA drives in the array powered down most of the time, providing significant power and heat load savings for a data center.

Data movement between tiers in an active archive can be accomplished via administrative action (manually moving files to different storage locations) but this is a cumbersome process and can result in data loss. To alleviate this issue, most organizations that use active archiving rely on products that automate data movement between tiers. These products often also include the ability to protect content by cloning of files and allowing them to be placed in a cold archive. Active archiving products are built to include or work side by side with content asset managers.

Some of the companies making software used for asset management and digital archive management in the entertainment and professional media markets (either by themselves or in combination) include: Dalet, Etere, Front Porch Digital, Masstech Group, Quantum, SGI, SGL and XenData. Cold archiving is done by companies such as Iron Mountain that specialize in retaining records and assets in controlled environments.

Issues and Opportunities in EM Archiving

In this section we shall examine some important issues in the archiving of digital entertainment and professional media content in order to get a better understanding of them and to understand how they can be managed. This will include what is involved in dealing with format conversions, protection of archived data from disasters at the primary data center, longevity of various possible storage media and total operating costs in maintaining a digital media archive.

Format Conversion and Long-Term Archive Management

Over the course of time storage media formats and interfaces become obsolete. This poses special problems for a long-term digital archive. In the case of retained source media extra copies of the reader devices for the media as well as systems that these readers can attach to are required to access these media in the future.

If the retention time is especially long this could even include a string of devices that can be used sequentially to convert older media contents into then current formats and with interfaces that then current devices support.

Thus to be secure in the future use of source media an archivist needs to also keep all the hardware and other software needed to read and convert that content into a form that will be readable to future hardware and software. Over the course of time this can become quite a collection of devices and can become a serious management issue itself.

Regardless of whether the original source media are retained if the content data is transferred into a media archive system using hard disk drives, optical disks or magnetic tape format obsolescence still forces the administrator to convert older storage formats into newer ones as time goes by. Storage formats that include backwards compatibility mitigate this issue, but do not solve it.

This issue becomes more pronounced as the size of the digital archive increases. As the size of an archive grows the time to transfer or convert digital content from the old media to the new can be overwhelming unless there are sufficient old and new resources on hand.

In a sense, archiving (even cold archiving) should not be a static process. When the archive load becomes too large choices will have to be made about which content to transfer and preserve on the new format. Format choices should always be made with a consideration of backwards compatibility. Otherwise archive transfers could become a constant process.

The rate of obsolescence for storage device formats varies with the devices. For hard disk drives the stability of the interface and the drive specification that controls that interface sets the backward compatibility. For SCSI drives and for fiber channel disk drives that use the SCSI specification there is very long format stability (backward compatibility) on drive commands.

On the other hand the SCSI hard drive connection interface evolves to a faster version, roughly every 5–7 years. Recently both SCSI and ATA, found in the past in personal computers, made a conversion from parallel to serial connections. The resulting SAS and SATA interfaces are completely different from the interfaces used in older format drives. Thus disk drive interfaces change relatively frequently with time from an archivist point of view.

An archival system built around the use of disk drives must take this into account and either have a means to migrate to newer drives using adapters as the old drives wear out or eventually to move data from the older drive arrays to a new set. Thus the useful life of an active disk array is probably somewhere in the range of 5–7 years (the actual functional life with appropriate spares could be as high as 10 years). This time frame will also be influenced by service contract costs. After 3–5 years service costs for arrays tend to go up dramatically resulting in the desire to swap an array, even if it is functioning adequately.

For a MAID system with less active disks, this time period could be longer than 10 years. Often disk storage systems (as well as tape and optical) are replaced not because they no longer work but because there is technology with much more storage capacity, better performance and lower operating costs available.

The format obsolescence rate for tape (a common digital archival media) is variable depending on the format and technology. Different tape standards take different approaches.

For example, the LTO Consortium's roadmap communicates the member's intention to provide read/write capability one generation back and read capability for two generations back. The intent is to provide that capability at each specific generation's original density and performance. Thus an LTO Gen4 drive which is capable of recording 800 GB per cartridge at 120 MB/s must be able to read and write up to 400 GB on Gen3 cartridges at the manufacturer's originally specified Gen3 performance and must also be able to read Gen2 cartridges, again at the originally specified performance.

The LTO Consortium has a track record of introducing new generations every 20–30 months (note that DLT tape technology roadmaps have similar prior generation read requirements). Thus archival tape systems have a total format life (for reading) of about 6 years. An existing tape system could last much longer and the media is rated to last up to 20 years. Tape storage systems (especially libraries) probably have a useful life of 10–15 years.

Optical storage media may be more stable in media development than tape but generally there is a major format change every 10 years and interfaces develop like those of other computer peripherals. Optical storage media of good quality can probably last about 10 years under the right storage conditions and optical archive libraries probably have a useful life of about 10 years.

Thus an 8–10 year life of an archival storage system is likely for all these storage devices. Sometime before an older system is to be retired, its content must be moved to a replacement storage system.

To handle this format conversion, manual processes could be used, but they are error prone and time consuming, taking staff away from other projects. Storage management software should include a way to automatically control the format refresh. Management software can read archive data from an older generation tape environment, verify that the content is still intact and then automatically rewrite to the latest generation of tape.

By having this process follow rules defined by the client, their specific data retention and protection needs can easily be factored into the process. This also allows for maximum reuse of tape media and ultimately minimizes long-term data retention cost. Over a 100 year archive life this process is liable to happen at least ten times. This operation benefits from active software management to make sure such media progressions occur according to a pre-set schedule and that transfers are successful before the older media and storage systems are retired.

Creating Better Metadata for Archived Content

Creating better metadata to represent the data in the archive and creating the means to manage this metadata are keys to better search and discovery of historical

content. While today much metadata is entered manually, technologies are being developed to enable searching data based upon matches with audio, still or moving video content. As these technologies develop the resulting metadata could be incorporated into professional video metadata formats such as MXF to create powerful ways of accessing older content archives.

Storage devices and storage management software should be designed to make use of these automated metadata generation capabilities as they are developed. Archive management software should search through active archive content, creating new metadata based upon rich media attributes. In the absence of this, or in environments that use customized metadata formats, archive management software should be as tightly aligned with the asset manager as possible so that artists can track and manage content as they need.

Decreasing the Overall Costs of Archiving

The cost of archiving is a function of several factors. Some of the more important of these factors are storage media utilization and the costs of maintaining the data center where the storage assets are kept. Storage management software is a key element of reducing the overall costs of storage.

Storage management that can find unneeded duplication of data and not back up multiple copies of the same files is one method to reduce overall system digital storage capacity as well as reducing the bandwidth requirements for moving this data around.

Although generally higher in archive environments, storage capacity utilization is usually rather low. Disk storage utilization of less than 50% is not uncommon in many data centers. It is possible to improve storage utilization by creating virtualization of storage assets so that storage can be provided to the user as needed rather than relying on pre-provisioning storage. By reducing the overall hardware and facilities costs (which tend to be much greater over the long run) good storage management software can more than pay for itself.

The cost of maintaining storage facilities greatly exceeds the initial hardware costs. Heating and air conditioning are essential to successful operation of an active storage facility. Reducing the heat load from the storage systems will do a lot to reduce overall facility costs. The use of smaller disk drives in drive arrays (particularly if these are lower RPM SATA drives) will reduce the heat generated by these drives and as the drives can be packed tighter in an array than larger form factor disk drives, the overall storage array footprint can be smaller for an equivalent storage capacity. This reduces the size of the data center space required for the active archive.

For larger archives tape is an important format because it does not require a large amount of power. Media is offline except when in use, in a tape drive. The tape library itself requires no power except for the robotics and drives, which when not in use are idle. Tape libraries can scale to multiple petabytes of storage

without requiring significant power consumption in comparison to a disk array of the same capacity.

Projections for Various Storage Media Used in EM Archiving

Table 11.2 gives projections for annual demand of new storage media (tape, optical or hard disk drive) for entertainment and professional media archiving of newly generated content. Digital conversion of analog content and its preservation can increase these numbers considerably but prediction of this trend is hard to do. Suffice it to say that the overall demand could easily be twice that shown here.

As can be seen in this table tape remains the most used archival storage media. The anticipated growth in optical media depends upon the development of advanced high capacity optical technology such as holographic media.

A constant problem with a cold digital archive is staying ahead of the obsolescence of the storage media used. As time goes on digital storage technology improves and the storage media and drives change. If the data storage on a cold archived media is not migrated to new storage technology when it becomes

Table 11.2 Projections for the use of various archival storage media in the media and entertainment industry

	2005	2006	2007	2008	2009	2010	2011	2012
Archiving new content (TB)	61,571	82,820	111,243	153,291	201,300	253,149	310,528	379,518
Capacity change (TB)	16,436	21,250	28,423	42,048	48,008	51,850	57,379	68,990
Digital tape (%)	88.0	86.0	84.0	81.0	78.0	75.0	76.0	70.0
Optical disk (%)	11.0	12.0	13.0	15.0	17.0	19.0	21.0	23.0
3.5″ ATA HDD (%)	1.00	2.00	3.00	4.00	5.00	6.00	7.00	10.00
Estimated half inch tape cartridges	43,830	46,859	47,750	56,765	53,495	48,609	48,453	48,293
Estimated optical disk units	51,657	54,255	57,734	66,392	58,296	54,730	54,771	61,030
Estimated 3.5″ ATA HDD units	548	1.062	1,705	2,803	3,429	3,889	4,463	6,899

available, there is a danger that the data will not be readable as time goes by. Thus successful preservation of archived digital content must also include data migration management to deal ahead of time with the risks of format obsolescence.

Broader Asset Management Systems

A final topic of discussion is how archiving fits into a larger asset management system. An asset management system is a set of processes designed to control digital assets used in active and inactive projects. Processes should cover every aspect of the content lifecycle from tracking metadata, physical location from active usage to archive, and handling data protection.

While there is no one catch-all product that covers asset management, some broad strokes can be made about the design of a system. A digital media archive management system should do the following:

- Protect the digital assets for many years through active management and checking of the hardware and storage environment.
- Perform systematic format updates and transfers of the digital content as needed to avoid potential access loss due to format obsolescence.
- Alert the administrator of problems with the content, environment or storage hardware early enough for the administrator to take action and if necessary shut down and protect assets in the event of a crisis.
- Provide organized access to the digital archives by retaining an appropriate database of content metadata and perhaps indexing of the actual content on those assets.
- In case of an active archive, control the access and delivery of digital content when needed and in a timely manner.
- In case of a cold archive, initiate delivery and mounting of the media from the cold archive and delivery of the content where it is needed.
- Manage disaster recovery requirements such as mirroring or backup of managed data to remote data centers to make sure content can be recovered if the original copy is somehow damaged.

New developments in long-term archive management include:

- Active indexing of still and moving digital images during the archiving process – more advanced metadata creation and management allowing easier search and use of this content.
- More active and continuous checking of archived content to make sure the data integrity is OK and to detect growing problems.
- File-based access to digital content from the storage media to enhance content access.

Conclusions

This report has explored the important trends and drivers for the archiving of entertainment and professional media content. Management software as well as hardware play an important role in digital media archives. We have looked at the requirements as well as future expected developments in storage management. We have also looked at some of the providers of digital media management software.

The decreasing cost of digital storage and the capture and editing of content enabled by digital technology has increased the number of facilities and producers of content.

Digital technology and increased available communication bandwidth have also allowed the development of new ways to distribute content such as the internet and mobile phone networks. This has also increased the supply of new professional (as well as non-professional) content. Higher resolution is being required in professional content and the amount of digital footage being acquired for the final produced product has increased considerably. In addition, older analog content is being digitized in greater amounts with time. All of these factors have caused digital storage capacity requirements to swell.

As total content capacity increases so does the amount of content that is being archived in either cold or active archives (or both). Long-term preservation of large digital archives will lead the industry to solve new issues associated with format conversion, metadata creation and management as well as methods to reduce the total cost of operating a digital media archive. Preservation of digital data from the possible destruction of the primary storage system can be dealt with by having off-site copies of the content or by maintaining remote mirrors or backup of the content.