Content Lifecycle Management for Digital Content Archives

By Tom Inglefield

The storage industry has overwhelmed a number of industries with the current buzzword, "information lifecycle management" (ILM). This is a basic concept with a basic premise: "Use the right storage media for information during its business life, to meet business creation, usage, and regulatory requirements of the information, at a cost consistent with its current usage value to the corporation." In short, when you create a piece of information/data, use the proper storage device and media during its life; that is, fast, more expensive devices (generally disk) during times of high recall and use, and slower, very inexpensive devices during periods of low recall and use.

ILM draws from concepts and architectures in information technology and data processing systems and applications. Many ILM messages are heavy on copy 1 (Copy 1 is the primary, currently used version of the information or content.) and management of backup copies. Many ILM concepts are directly related to the technologies delivered by the vendor (For example, disk vendors talk about copy 1 or 2 (copy 2 through n are defined as backup copies of the primary working copy, copy 1), network vendors talk about all information being in the network, and so forth.). However, for content producing, delivery, or distribution companies, a new ILM is required for content lifecycle management (CLM). Why? The workflow is different; content is truly the entertainment industry's asset. Figure 1 shows the relationship between content life, content use, amount of content, and possible storage device options. As content moves through the work process, many copies are created as the material is refined, edited, animated, etc. Some versions of the content will continue through the work process, while other versions should be archived for limited future use. At the end, content or programming will be distributed in its final form for consumers.

By applying the precepts of information lifecycle management to the entertainment industry there are many examples to describe this concept. For example, the business cycle for news content has a variety of usage characteristics. Today's news demands high-speed, highly reliable storage devices that integrate with today's digital news rooms to bring stories to air during each newscast. As days pass, news directors may decide to save only finished stories, keep minimal story-related material, or keep all material for long periods of time. One can easily argue that with properly costed storage devices and archives, much of the "wild footage" would be retained indefinitely. The value of "wild" or "raw"

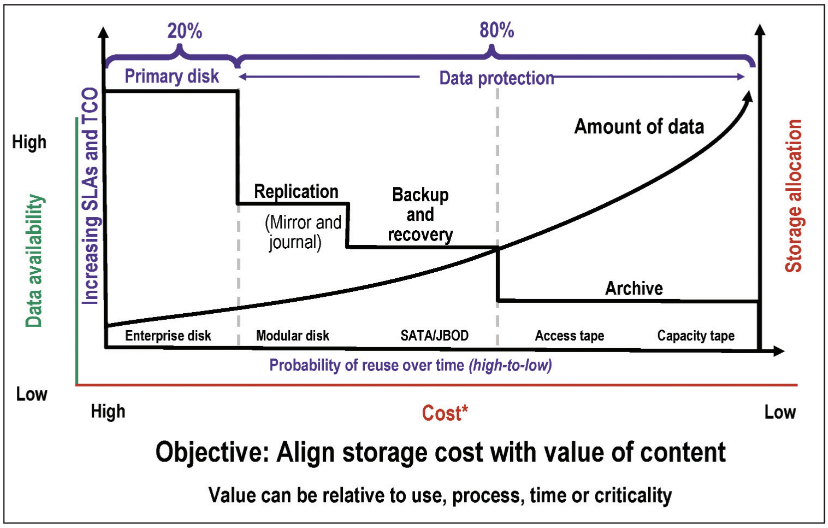

Figure 1. Content value and storage.

material may not be realized until a story line is picked up for historical reference when it may be invaluable for story continuation. Applying a CLM methodology would place today's news on high-performance disk devices (possibly RAID), last week's news on highly available, high capacity (less expensive) disk devices (possibly SATA), and the long-term news stories would be (stored months or years) on automated tape devices.

This paper examines digital content archives and applies the concepts of information lifecycle management. However, information lifecycle management concepts were originally written for information technology and data processing systems and applications. ILM does not necessarily address the concerns of an industry in which content is the key business asset. Content lifecycle management addresses the business, content, technologies, and applications for the entertainment industry, including filmmaking, program creation, program distribution, program playout (broadcasting or video-on-demand), etc.

Content Management Workflow

Figure 2 presents a content workflow that can be used for many different business operations in broadcast and entertainment. The workflow demonstrates typical play-to-air or transmission steps from ingest (step 1) to scheduling or automation, to playout (step 3). This is a fairly simple process that, in many cases, still relies on manual retrieval of program material, ingest, and then playout. Deploying a digital content archive can provide dramatic improvement to this process by using an automation system to move content to archive and then recall content from archive for play to air. Content moving into and out of the archive moves at digital speed of the storage devices, rather than in realtime. Taking advantage of digital-speed versus realtime is very important to workflow improvements. With high-speed disk storage devices, content starts returning from the archive with no delays, at many times faster than realtime. However, storing content on high-speed disk is more expensive. In the worst case, taking advantage of very inexpensive data tape storage, approximately 1 minute after request, the content will begin moving at approximately 240 Mbits/sec.

Media asset management (MAM) (step 2) deals with a key element of an archive, logging the content for later retrieval. Logging and indexing are dependent on the business requirement for use of the content. The metadata is typically cataloged in the MAM database managed on disk storage for very fast retrieval and review.

Common questions include:

- Is low-resolution browse required?
- Is key-word or topic search required?
- How granular is the search requirement (scene, topic, date)?
- How granular is the recall requirement (frame accurate, time code, etc.)?
- Is script search required?
- How is the content to be logged?

During ingest and logging, several things happen to the content. The high-resolution material can be immediately sent to archive for later use in the work process. This leverages the cost-effectiveness of storage media. The metadata collected and low-browse can be inserted into a metadata catalog for normal everyday use by people browsing the content. The metadata is typically cataloged in the MAM database managed on disk storage for very fast retrieval and review. This accomplishes several things:

- Low-resolution browse can be viewed using standard networks and web browsers.
- Metadata can be searched using standard networks and web browsers.
- Rough cuts and EDLs can be created, shared,

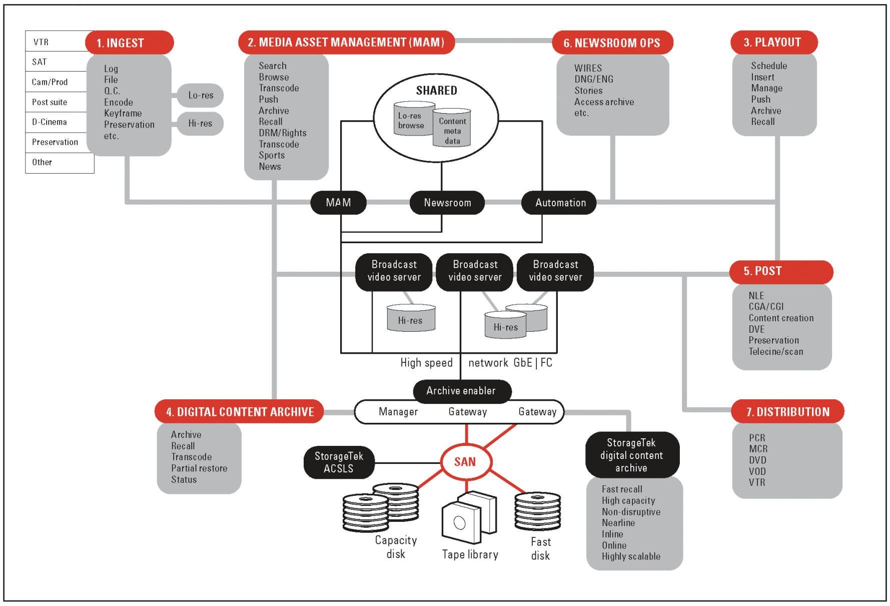

Figure 2. Digital content archive workflow.

reviewed, and approved, minimizing costs for expensive systems and applications.

• High-resolution copy can be extracted from the archive later in the workflow when needed.

• High-resolution copy can be extracted, based on frame-accurate time code reference, minimizing the movement of material through networks.

• High-resolution editing systems can better utilize expensive components and storage for realtime editing.

Other steps in the workflow are also enhanced during post-production work (step 5) and specialized applications such as news rooms (step 6) by using search capability to pull content from the archive for reuse, repurposing, and so on, as needed.

Consider the process of serial or episode creation. After the first episode is finished, all material can be logged and put into the archive, freeing up resources on the production system. If work is being done on the sixth episode and the director wants to include work from the first episode, it is very easy to browse the asset management system, select the material required, and pull the material from the archive.

In all steps, content can be cost-effectively managed in the workflow through better access to deeper stores of content that can be easily recalled for use. In addition, leveraging the different price breaks in different storage media results in improved cost and margin per title of content.

Content Archive Architecture

The architecture to support CLM, digital content archive, and workflow must meet many requirements from the technologies already existing in the content plant. For example, most video servers contain proprietary storage sold by the server vendor. If more storage is needed, it must be purchased from the server vendor, a rather expensive proposition. However, if the content is to be shared among technologies such as playout servers and nonlinear editing systems, each system must contain its own storage and material passed through a network (or tape-based "sneaker-net"). Implementing a content archive opens the storage platform for sharing material among the various technologies.

To enable and support content storage and movement requirements, content archives are typically built as open shared storage platforms, using storage area networks (SANs). This provides several benefits, including:

• Flexibility to add storage devices to meet volume and throughput requirements.

• Adaptability to meet changing workflow requirements using a variety of storage devices.

• Flexibility to change and add new storage media as they are introduced.

• Flexibility to handle a variety of storage devices and media.

• Ability to scale performance characteristics of the archive, based on recall rates and bit rate of the content.

To achieve all the goals and business requirements of the content workflow, the content archive can be built with a variety of storage devices, including:

• High-speed fiber channel RAID devices for metadata and caching high-resolution content. This provides very fast access to content and fast streaming of content back into the work process, at a premium price. In addition, this provides very fast access to a large number of metadata process requests for content recall.

• High-capacity SATA RAID devices often used for high-resolution content. This provides fast access to very large high-resolution content files at a much more affordable cost. However, many SATA systems are not designed for a high number of metadata processing requests. These devices are best suited for a moderate number of streaming recall requests.

• High-capacity automated data tape libraries for very large volumes of high-resolution content. This provides the most inexpensive storage media for very large high-resolution content files. This storage media is several orders of magnitude less expensive than disk-based storage. However, there is a delay to start streaming the content from the digital data tape to the requesting process, generally less than one minute. Once moving the content from digital data tape, content can generally be streamed many times faster than realtime.

In addition, digital data tape is a widely known and highly used storage media with a long history of reliability and longevity.

As media asset management systems are introduced, most already use databases and the concept of open shared storage. By architecting the content archive as an open shared storage platform, storage can be carved out and shared for specific MAM applications.

The "glue" that makes it possible to implement this architecture is the content archive manager, or shared file system. These software enablers accomplish several things, they:

• Provide integrated program interfaces to the proprietary interfaces used by the various broadcast plant technologies. These interfaces may be implemented as APIs, FTP, file system protocols, or drag-and-drop clients.

• Provide streaming capabilities to move the large content files into and out of the main content workflow and technologies, i.e., video servers, NLEs, newsroom systems, etc.

• Automatically, manage a number of storage devices and storage media to maximize throughput and minimize long-term cost of content storage. Content management on storage is based on business policy and rules set-up in the archive management system.

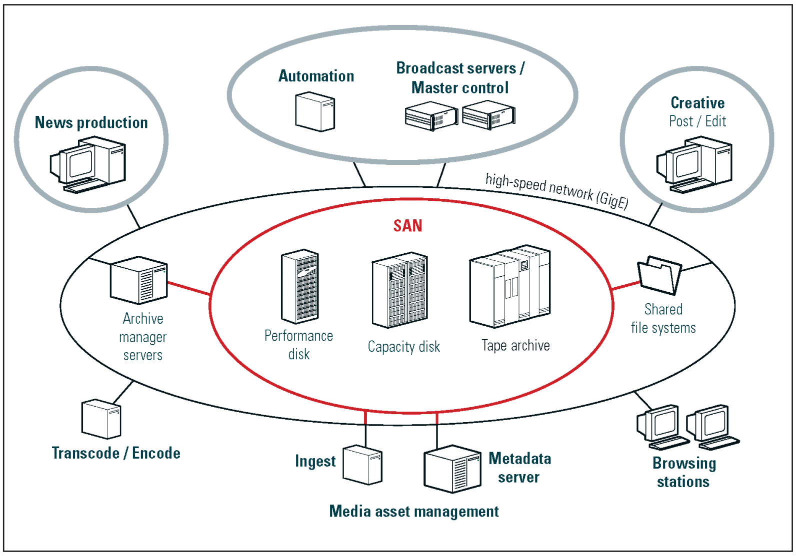

Figure 3. Digital content archive architectural overview.

The architecture (1) enables use of standard storage products, leveraging storage price points (2) integrates content archive into the broadcast or entertainment plant and operation, and (3) enables workflow improvements based on simple access to archive (Fig. 3).

Content Lifecycle Management and the Numbers

There are many factors to consider when building the content archive that will affect the total cost of ownership and the return on investment. Each business must determine the workflow of content as it moves from creation through production to distribution and where the business benefits will be derived. In most cases, there will be multiple workflows in the plant, and each must be considered. Key considerations include:

• What is the creation process? How is content acquired? Can a format of the material be created and used at browse or low-resolution in the work process? How soon in the process can a browse copy be used, and can the high-resolution version be moved off expensive storage?

• What is the production process? What is the ratio of work product versus finished product? How soon can work product be moved out of the production schedule to archive? How often will work product be recalled for further edits?

• What is the distribution process? When and how often is the finished product used? Does the finished, high-resolution version need to be instantly available all

Table 1—Typical Storage Costs

Server-based disk (high-speed fiber channel raid)	>$50/GB
Open systems disk (high-speed fiber channel raid)	$18-25/GB
SATA disks (high-speed fiber channel raid)	$4-10/GB
Fast access data tape	$2-4/GB
Capacity data tape	$1-2/GB

the time? What formats must be kept for distribution?

These considerations must be evaluated within the context and confines of the following facts and guidelines:

- Finished feature film, 2K resolution = 2TB (terabytes).
- Generated work product for feature film, 2K resolution = 10-100TB.
- Finished feature film, 4K resolution = 8TB.
- Generated work product for feature film, 4K resolution = 40-400TB.
- Finished high-definition programming at 50 Mbits/sec = 23 Gbytes/hr.

Content lifecycle management evaluates content volume (content use over time with storage costs). Typical storage costs for high-performance disks are in the $2,000 to $4,000-per-terabyte range. In contrast, storing content on data tape can be much less (as little as $400 per terabyte). For example, data tape media can store approximately 9 hours of 50 Mbit/sec video for less than $10/hr. With high-definition content and feature films consuming large amounts of storage, it is easy to see the possible savings by using different storage media during the work process (Table 1).

Once a cost estimate for building an archive has been determined, a number of other costs factors need to be evaluated in the work process, including tape machine maintenance costs and down time, replacing content (if possible), eliminating dubbing and redubbing masters, labor costs, moving content and tapes through production, review and approval, and real estate savings, due to the density of digital storage.

Conclusion

Content lifecycle management is significantly different from ILM. Although both have a basic premise of storing content (data) on the right piece of storage at the right time at the right price. In ILM, data is not the base product produced, rather the information about products and services produced. ILM strategies are developed to meet business continuance, disaster recovery, and regulatory issues. While important, these are not core to the business.

Content is core to the entertainment and broadcast business; it is the product. CLM is more concerned about the business process and workflow to create, produce, and distribute content. With CLM, content is the product, and it is critical to understand the content, business process, and workflow. Once these requirements are well defined, storage architectures and storage media can be applied for long-term management of the content, while maximizing content storage in the process and minimizing storage costs. At the end, more content can be stored in digital format over a longer period of time for much less costs.

Presented at the 146th Technical Conference and Exhibition in Pasadena, CA, October 20-23, 2004.

THE AUTHOR

Tom Inglefield is a specialist at StorageTek, focused on media and entertainment content workflows, architectures, and storage. He works with studios, broadcasters, post houses, and others to develop digital content storage and archives. He architects storage solutions and offerings that solve very large video content library and archive problems and issues. Inglefield began his StorageTek career in 1982 and has developed several industry-leading solution offerings working with storage providers and application partners.

Inglefield has served as the director of product development for a streaming media startup. Prior to that he was chief architect responsible for developing storage software, hardware, and service products for the very large database/data warehouse market. Inglefield has extensive experience in marketing, storage technologies, business development, product development, and database technologies.