

Information Lifecycle Management for Broadcasters

THOMAS E. HALLEWELL

Radio Free Asia
Washington, DC, USA

ABSTRACT

With today's increasing hunger for "original" media programming, there is an expanding market for "vintage" content, as well. It is impossible to know which of your archive programs may be valuable assets in the future. As broadcast "converges" from analog to digital, and the cost of massive-capacity storage goes down, it becomes tempting to save all your data and programming archives. But what should be saved, how should it be saved, and for how long?

The first step is to define the types of data you are storing and decide how long you should keep each data type. A typical broadcast organization will have the following types of data:

- Corporate Data
- User Data
- Program archives
- Production outtakes

Your organization must determine a policy for how long each of these data types should be stored based on the following factors:

- Volatility (how long will the information be relevant)
- Accessibility (how quickly do you need to be able to access the data)
- Sensitivity (how private is the data)

This paper hopes to help get your organization thinking about information lifecycles so that you do not fill up all your storage-or lose valuable content and data.

WHY MANAGE DATA STORAGE?

For the purposes of this paper, we will use the terms data and information interchangeably to refer to any digitally stored information, including e-mail, databases, playlists, memos and program archives. As you move toward Information Lifecycle Management (ILM), it is necessary to start thinking of data as "information blocks" as opposed to the traditional terminology of "data blocks." This will help us further down the road when the storage devices change format and data is no longer stored in "blocks" on the media.

It is no easy task to completely rethink your storage strategy and infrastructure. Because implementing a comprehensive storage policy is labor and cost intensive, it may be difficult to sell to upper management. Plan to budget around five dollars on ILM implementation for every dollar spent on storage. Although it is a daunting task, it is enormously beneficial to develop a storage hierarchy that is relevant to your specific business and company. As your company grows, so will its storage requirements. According to a recent studyⁱ, the amount of new information stored doubled between 2000 and 2003. 92% of that data is stored on magnetic media, mainly hard disks. If this trend continues, storage will become a significant part of your company's overhead. Using it wisely will not only reduce your bottom line, it will also help you access potentially valuable content more effectively.

Legal Compliance-A number of laws and regulations have been enacted in the US, requiring certain data to be stored for long periods of time, but not many of them have much impact on the broadcast industry. If your company is publicly traded, Sarbanes-Oxley may be relevant, but only in regards to your financial data. Most likely, most of the burden of complying with S-OX will fall on your accounting company. That said, it is virtually certain the archiving of digital data and transactions will become more and more a part of the business landscape, so you might as well start planning for it.

Content Management-As broadcasters, our business revolves around content. To a greater or lesser degree, we record and document the life of our community, our nation and our culture. The events we broadcast, be they contests or guerilla skirmishes, are snapshots of a singular moment in time and are irreplaceable. Many broadcast organizations, such as BBC, have found ways to capitalize on this unique intellectual property in a big way. Although you may not have footage of The Battle of Britain, your company still needs to position itself to be able to store, access and get value from its historical content. Programming based on historical material is cheaper to produce than a brand new program. There is also an appeal to "vintage" advertising and images. You may possess footage of a public figure before he or she became famous. To find that footage on a "just in time" basis

is the goal of any content management system. Your storage management is a key part of that process.

Manage Corporate and Financial Data-Businesses generate more and more data as they age. Much of this data, such as policies and reports, is redundant or obsolete. To prevent duplication of data and the distribution of incorrect information, it is important that historical data be kept separate from current data. Information Lifecycle Management (ILM) can assist in this process.

Disaster Recovery-Every organization needs a way to prevent their data from being destroyed in a disaster such as a fire or system crash. With analog media such as tape and film, it was possible to mitigate a lot of damage with fireproof and waterproof containers. This is less practical with data. A sound backup and archive strategy can help minimize your data losses and reduce your downtime.

Security-Your company may wish to limit access to certain confidential or privileged information. As data ages, its sensitivity is often forgotten and confidential information is stored in a non-secure manner. It is possible to tag files with a confidentiality rating as they are created, so that you can flag this data for special treatment.

Expiration Dates-You can also tag data with a target deletion date so that you can automatically purge data as it loses relevance. This will lower your storage requirements, thereby reducing the total cost of ownership of your company's storage.

INFORMATION LIFECYCLE MANAGEMENT

The last decade has seen a profound change in the production and archiving of broadcast material. We have gone from hand-splicing big tapes to digitally editing smaller tapes to editing and storing directly to computer hard drive. Storage capacity has increased exponentially, while storage price has dropped in reverse proportion, making it possible to store "everything" "forever".

Assuming you even wanted to do this, how would you go about it? And how would you find this data when you needed it? Information Lifecycle Management is a concept that has been bandied about quite a bit of late. While a certain amount of this may be vendor hype, a certain amount is not, and the basic model could be extremely valuable to broadcast organizations, whose capital, more so than any other business, is in their tapes.

Information Lifecycle Management is a relatively new concept in Information Technology, and as such, is still very much under development. The standards and

solutions are still somewhat undefined. Hopefully this paper will get you started thinking about how you might be able to apply some of this technology to your broadcast operations and help you organize your company's data assets so that valuable material is not lost due to bad management.

Although it's a fancy new word, most of the concepts of Information Lifecycle Management are simple. As its name implies, Information Lifecycle Management is less a new technology than it is a data-centric way of thinking about your workflow and business processes. Properly implemented, ILM can help you base your storage of, and access to, data on your real business needs, and enable you to create flexible strategies for archived data in the future.

Clearly, data is useful only if you can find and access it within an acceptable timeframe. If you have to search thousands or millions of files or tape cases to find the data you are looking for, the moment of its greatest relevance and value may have passed, or the cost of retrieving the data may be greater than its value.

STORAGE STRATEGIES

In order to maximize the return on your archives, you must first develop a strategy. What kinds of data does your organization create and what are your requirements for retrieving it?

One useful way to look at storage is as pools.ⁱⁱ There are three main types of pool; data type, data use and storage pool. Classify and match your data to each of these three pools to determine your storage strategy.

DATA POOLS

Data comes in three main types-structured, semistructured and unstructured.

- **Structured data** is sortable data such as the data in databases.
- **Semistructured data** is searchable text data such as spreadsheets, e-mail and word processing documents.
- **Unstructured data** is encoded data such as audio and video files, which can be sensed.

USE POOLS

Use is more nebulous and must be defined in relation to your organization. Some use categories could include frequency of access (active vs. inactive), update frequency (fixed vs. changeable), and sensitivity (confidential vs. public).

Any broadcast organization will presumably have corporate information, such as financial and personnel data. This data is volatile, sensitive and needs to be kept forever. You will also have corporate policies and procedures. These are rarely written to, but need to be kept available for frequent read and occasional write access. They, too, need to be kept indefinitely.

Content archives are the largest, and most valuable, information asset in a broadcast organization. Archival programming could be of immense value in the future. This corporate intellectual property needs to be searchable for future data mining and should be accessible on a read-only basis. Program and production data are unique to our industry in that they are generally written once and edited many times. Additionally, the size of this data is proportionally larger than a word-document transcription of the same information. This data set is our biggest challenge in the retrieval process because of the inability to search within the object, the way you would a text document. Current technology prevents the ability to do such a search, but that may not be a problem in the future

Finally, there is user data, such as notes, e-mail, memos, photos, etc. User data will probably require the most thought of any of your policy decisions for a number of reasons. Users value their “private” data. Although 90% of this data is of no business use whatsoever, users will complain the loudest if their personal data is lost. However, this data expires when the user leaves the company and can usually be deleted. User data must be kept confidential and in some cases can be subpoenaed for legal reasons. User data is very similar to corporate data, but has the unique value of personal ownership. This value, in the eyes of the user, puts user data at the forefront of importance.

STORAGE POOLS

Storage pools are the actual place in your storage system where a file will reside at any given time. The four types of storage are online, midline, near-line and offline.

Online Storage is high-cost, high availability, high speed hardware storage such as NAS or SAN. This is the best place for frequently modified and accessed structured data such as databases and recent

unstructured data such as pre-air programming that still needs to be edited and streamed to air.

Midline Storage is more affordable, lower-performance storage, such as IDE RAID or server-attached storage. Current unstructured and semi-structured data reside here as they prepare to be migrated to the next phase of storage.

Near-Line Storage is just coming of age. As writable disks such as DVD-R increase in size and lower in price, disk to disk backup will eventually replace disk to tape. Writable or rewritable disk libraries will enable affordable rotation of backups.

Offline storage will always be your last defense against disaster. A full backup of all current data and all relevant archives must reside offsite. With internet bandwidth becoming cheaper, it may be affordable to perform incremental backups to hard disk at branch locations during off-business hours. This can vastly reduce your total cost of offsite storage.

There are several conflicting dimensions to data storage.

- **Age vs. access:** As data ages, it needs to be write-accessed less and less. Some data still needs to be read; other data should never be read. Some data needs to be accessed in seconds; it may be acceptable to wait for days for other data. Think arrest warrant versus outstanding parking fines.
- **Replication/Redundancy vs. Storage Space:** You may not realize it, but some data is redundant by nature and to back it up would be redundant. We aim to keep all relevant data in one form and as “one copy.” Any more copies than needed are waste. Think customer information or email addresses that may be stored in multiple databases within the organization. Or 5-minute audio files that appear as single files and are replicated within an hour long on-air show; which version should you keep and why?
- **Permanence vs. Timestamp:** Some data needs to be kept for the life of your organization - and perhaps beyond, while other data has a shelf-life based on a date associated with the file. Last access date, creation date, or an embedded deletion date can all be used as criteria for timestamp-based storage. There can be any number of rationales to justify a specific archive period: regulatory mandate, organizational needs, possible historic value, etc. It may be that your digital archives will

eventually be handed over to an institution such as a library or university.

PRESERVING ARCHIVES

Information Lifecycles also involve two very concrete lifespan issues -the lifespan of the medium on which the file is stored, for example, a DVD, and the eventual obsolescence of the format. Digital formats age as by no means etched in stone, they are often superseded by new formats, particularly if they are proprietary (think 5.25 inch floppy drives).

DATA LIFECYCLES

All data has a lifecycle we can measure and predict. Initially, it is quite unstable - it is created, accessed and modified frequently; but eventually, every piece of data becomes relatively static and can safely be migrated to a more archival, less-editable medium.

It is important to recognize different types of data have different lifecycles. Corporate data such as financial data or policies tends to be modified frequently early in its life, then mostly read later on. Content is generally written, edited down, and saved. Both the raw content and the edited program should be archived and indexed if possible.

How long do you want to keep data? The oldest pieces of “analog information” produced by humans are probably cave paintings and are at most 40,000 years old. The oldest pieces of digital information are, at most, as old as the medium and thus are perhaps 70 years old. Seen in this context, the term “forever” takes on a very different meaning. “Indefinite” might be a better way to phrase this goal.

MEDIA LIFECYCLES

Something that is rarely considered when developing a storage plan is the lifespan of the storage media itself. Currently, even non-volatile storage is very impermanent. Think of the thousands of hours of vintage films that are decaying right now, requiring thousands of dollars per hour to preserve.

As digital media ages, this will become a huge problem. While manufacturers claim lifespans of 50 to 100 years, the National Institute of Standards calculates that, handled properly, a DVD or CD can last up to 30 yearsⁱⁱⁱ. The average lifespan of a hard disk is approximately 5-15 years.

It is vital to keep your data in readable formats on fresh media. Plan to convert all the files in your “permanent” archive to the latest format and copy them to fresh media every ten years at most. This process is known as refreshing.

FORMAT LIFECYCLES

In 1086 William the Conqueror commissioned the Domesday Book. It is still available to be read by qualified researchers in the British Public Record Office. In 1986 the BBC created a new Domesday Book at a cost of £2.5 million. It consisted of 25,000 maps, 50,000 pictures, 60 minutes of video, and millions of words, but it was made on disks which could only be read on a special computer, all of which have been either decommissioned or are no longer functional. This much-touted “Domesday Book V2.0” was viable for less than 16 years.^{iv}

So another important factor to consider is the rapid evolution of file formats. Somehow, you have to ensure your digital archives will be readable or playable when someone needs to access them. There are two general approaches to this-migration and emulation.

Migration-Migration means periodically moving files from one file encoding format to another to keep up with state-of-the-art formats. (An example would be moving a Wordstar file to WordPerfect, then to Word 3.0, then to Word 5.0, then to Word 97.) Migration mitigates file compatibility issues by gradually updating your data to a limited number of contemporary file formats.

Emulation-Emulation focuses on the applications rather than the data, seeking to reproduce the computing environment of legacy software. So to use the example above, you would work on preserving a DOS environment so Wordstar could continue to be run, allowing you to access documents created in Wordstar. Over time, this tends to generate a more overhead as it becomes exponentially more difficult to support peripheral functions such as printing, duplication, etc., particularly as platforms themselves reach obsolescence. It could prove very tricky to migrate a file created on an Amiga to a Mac, particularly if the Amiga does not support networking.

While there is no easy answer to format obsolescence issues, it is something you must allow for any long-term archive plan. Since you will have to refresh your data periodically to keep up with your media lifecycle, this may be a good opportunity to update your file formats as well. We recommend you refresh hard disk storage at least every 5 years, tape every 10 and optical storage every 15. And put a plaintext ASCII file on the medium with as much information about your file formats and any special coding or behavior as possible. This will help with migration or emulation.

AUTOMATING ARCHIVES

One aspect which sets Information Lifecycle Management apart from past archiving models is that automation of the process is a central objective. There are a number of different ways to do this, and if you have some scripting background, it doesn't have to cost a lot of money.

You want to keep the location of your data transparent to your end users—they should be able to search and retrieve files, regardless of the storage medium or where the files physically reside.

Database backend- This is the most elegant solution, but also the most costly, both in terms of finance and labor. You create a database with fields for storage strategy and timestamps. One issue to remember is you will also have to keep the database upgraded and compatible with whatever automation scripts you have deployed with it. The advantage is you can have all the information about your data – creator, content, search keywords, shelf-life, purpose - in one place.

Embedded data- Most file formats now support the inclusion of user-defined fields in the header information. You could define some storage-based fields such as delete date and target retrieval time in the header and create scripts to automate archiving based on those values. For example, the AES standard WAVE file header format^v has several fields of interest, including Start Date and Time, End Date and Time, Producer Application and version ID and URL. There is also a 64-character User-Defined field which could, for instance, contain a sensitivity level and expiration date. There is also a TagText field which allows for descriptive text that could be used as search keywords.

Naming conventions- This is the poor man's solution. Create directories based on your archiving strategy and data types. For example, forever fast access, ten years medium access and 5 years fast, forever slow access. Then create subdirectories named based on the month of creation. Name your files with the creation dates in them, for example, nyc-interviews-2001-0911-0900.ram. This will enable you to create scripts to move and manipulate files based on their dates. For example, you could run a script every year to report all files more than five years old. Based on the report, you could move all "5 years fast, forever slow access" files from onsite, online storage to offsite, offline storage.

CONCLUSION

New strategies and technologies for storage are emerging, and they will revolutionize the way you

store your broadcast archives. Now is the time to start evaluating your data and its place in your enterprise. Categorize it, determine policies and start storing!

ⁱ How Much Information 2003, University of California-Berkeley Study, October 2003

ⁱⁱ Peter Kastner, A Six-Step Recipe for Adopting ILM Over Time, ComputerWorld, August 18, 2004

ⁱⁱⁱ Care and Handling of CDs and DVDs-A Guide for Librarians and Archivists, National Institute of Standards Special Publication 500-252, Washington DC, October, 2003

^{iv} Julian Jackson, Digital Longevity: the lifespan of digital files, Digital Preservation Coalition 2002

^v AES46-2002 AES Standard for Network and File Transfer of Audio - Audio-file transfer and exchange – Radio Traffic Audio delivery extension to the broadcast-wave-file format, Audio Engineering Society, 2002A