Archiving Movies In a Digital World

Dave Cavena
Chris Wood
Jeff Bonwick
Guy Steele
Michael Selway

January 10, 2007
VERSION 2.1, June 8, 2007

Table of Contents

Executive Overview

Filmed entertainment, the most visible of the Intellectual Property industries, quickly is moving to digital files from analog film for capturing and displaying images. When completed, this move will catch-up to the decade-old move to digital for intermediates and editing.

Archiving of this invaluable, irreplaceable content, however, still is done by outputting the images to film and storing the film in a vault under controlled environmental conditions, a process little changed by technology over a century and unyielding to current repurposing business models and market pressures.

Movie archiving, however, is coming under increased scrutiny as analog film quickly is being replaced by digital bits in all phases of the imaging process, from capture to post-production, to projection in the theater, to distribution to the consumer over the internet, in packaged media or via digital television, whether broadcast, satellite or cable.

Archiving for long-term preservation remains the only part of the workflow still reliant on film.

Using current technology and defined processes, recorded images and sound now can be archived digitally at a cost measurably lower than that of archiving film. With a faster recall from a Digital Content Archive this more responsive infrastructure can be used to create a more viable on-going market for content repurposing and re-monetization.

In this paper a viable, scalable, cost-effective model is presented for using digital technology to archive filmed entertainment for the centuries-long duration now expected from, and achieved with film. This model is based on Commercial Off-The-Shelf (COTS) technology; no special application software or systems are required.

Acceptance of this paradigm shift will not come overnight. The hope in presenting this paper is to initiate serious discussion, examine and improve the model, and assist content creators in creating further revenue opportunities for the ever-more expensive content they create.

Introduction

Feature movies, television broadcasts and other image and audio content are archived for two reasons.

Firstly, the content owner has a fiduciary responsibility to retain the assets with which he or she has been entrusted, making additional profit from them, if possible.

Secondly (though the primacy changes with the role of the person addressing the issue), movies are archived because they are the irreplaceable stories of cultures, civilizations, of our world. They are the recorded traditions of an age told by many of the best artists of their time or any other. They should not be lost to posterity.

With the average cost of a feature film approaching $100M, the desire increases to repurpose that content over longer time periods and larger markets. The current archive model tied to analog film does not readily support the efficient repurposing of content more than a few years from its release date.

Film archives of these irreplaceable cultural artifacts and revenue vehicles do not enable efficient access and re-use of these stories, nor an easy access to our past, whether the motive is profit or understanding. Archives must provide this access, or we are saving these irreplaceable assets solely for the sake of saving them and not for future generations to see, learn from and understand.

As noted by Mr. Daniel Rosen, Vice President, WB Technical Operations,

> Traditional analog methods of conservation and archiving cannot keep up with the growing demand for content in digital form nor the time-related degradation of the elements nor the rapid loss of analog process expertise... The creation of [digital] motion image content ...h as caused many motion image archives to look for hyperefficient ways in which to preserve digital data with digitally augmented analog techniques.[1]

A Digital Content Archive can allow that conservation and access, and can do so by providing infinitely-renewable, pristine copies of those stories efficiently in time, space and cost. A properly stored and migrated digital copy provides future generations with the view into our collective past that is required for them to be informed and understanding peoples, and provides content owners with long-term returns on the increasingly expensive stories being told through the medium of cinema and television. (Note: This paper deals with Feature movies only; the model for television and other recorded entertainment is similar, only the financials would change due to the increased amounts of content archived.)

The content stored in the vaults of the movie industry is valuable but very difficult to access.

How valuable is Hollywood's content?

> The copyright industries - which make movies, TV programs, home video, books, music and computer software - comprise an awesome engine of growth that nourishes the national economy. Not only is intellectual property America's largest trade export, and arguably its most influential, the industry sector is creating jobs three times faster than the rest of the economy. This sector generates more than 5 percent of U.S. gross domestic product.(2)

In 2006 the domestic receipts for MPAA films were US $9.49B; the global receipts US $25.8B.(3)

That's how valuable.

This content retains value over time. In order to capture this value, however, the content must be archived (4) in such a manner that it is easily recalled, and, when recalled, it is usable for its intended purpose of capturing new revenue or telling a story.

It is the opinion of the authors, and of experts in the described technology spheres, that digital archiving of filmed entertainment has reached the maturity necessary to be put into general use.

Archiving Movies

Movies currently are archived on film. As the movie industry has matured over the past century-plus, it has tried and discarded over 100 different formats, media types and standards(5), finally standardizing on a 35mm film format. Older movies in non-standard, obscure or obsolete formats and media are being preserved and archived to 35mm film by the Academy of Motion Pictures Arts and Sciences, in a preservation effort remarkable in its scope and effectiveness.

The standardization of 35mm film is such that nearly any theatrical movie recorded on 35mm, at any time, properly archived, can be retrieved from that archive and played in nearly any theater projecting feature films in the world today in 2007.

As movies increasingly are captured digitally and projected digitally, some major questions arise: Will it be possible in 2107 to retrieve from a film-based archive any movie placed there in 2007, and play it out in any theater in the world? If the movie was archived as film, will there be any projectors on which to project it or scanners to scan it? Projectors and scanners can be rebuilt, they're just electromechanical devices, but is this the desired future of this very valuable industry?

Or should the current – and future – digital world be embraced fully by the movie archive industry now that the technology exists to do so?

A Digital Content Archive will ensure the ability in 2107 – or in 2207 and beyond – to see the stories placed there in 2007, and to learn and profit from them.

A Changing World

Although film capture still reigns in Hollywood, this is changing. As with the adoption of new technology in many industries, the changeover is moving in stutter-steps. Magnified by the influence of personalities in Hollywood as the opinions of the quality of the new systems are debated, the adoption curve is not smooth, but the transition to digital capture is accelerating.

Several factors are driving this acceleration. The advent of digital cameras suitable for the capture of features is causing more movies to be shot digitally. New generations of cinematographers and directors graduating from film schools are eager to use these new technologies. Shooting digitally can cost less. All intermediate work is done digitally. The recently-agreed-upon Digital Cinema specification is resulting in a quick up-tick in digital projection at theaters.

All of these mean that at some not-too-distant date the first and only time a movie will exist on film will be for archiving. This means, of course, that the archived film no longer will be the image and medium through which the Cinematographer and Director captured and told their story.

At the Joint Technical Symposium 2004, Mr. Leon Silverman in his opening remarks posed the following question:

> Will there be a time when we can be assured that a very high resolution scan and workflow can reproduce every nuance that is the film and thus the digital data itself be considered archival? And when this is practical, as I believe it will be, will we not need a way to consider the digital data, the digital elements themselves archival as we consider film today - after all the data is the master record?

We believe the answer to that question is, "Yes," and that the time for this consideration is now.

Capturing, Digitizing and Storing the Image

While the ability to keep digital bits pristine over time has been the subject of much discussion for decades, and digital technology and processes often are assumed not to have kept pace with the need to archive irreplaceable content, it is time to revisit those assumptions.

The assumed but unknown qualities of each movie in current film archives, and the time and monetary cost of repurposing a film from its archive, are such that one wonders at the business case for preserving these assets for commercial reuse.

As Ms. Vicki McCargar noted at AMIA's DAS2007 (Los Angeles, May 11, 2007), "Preservation is a religion, but really it needs to become a business proposition. Saving for the sake of saving won't get you too far… More and more [companies are] attaching a business case to archives."

With the maturation of the digital age, and the growth of the internet and mobile consumer devices on which to consume content, movie archives are natural places to go to create new profit opportunities.

The archive business case is made viable through the implementation of digital archiving and the infinite renewability of archived content that it offers.

To capture revenue efficiently from previously-recorded content requires an archive and a workflow amenable to future repurposing of that content. A Digital Content Archive can provide this at a cost lower than that of a conventional film-based archive, and do so more quickly. This archive can provide quick and easy access to new revenue flows, in a time-to-market model that works in our fast-paced world of ever-new-devices on which consumers want to consume content – and on which content owners want to monetize their assets.

Digital Preservation Technology

Technology

Both hardware (computers, storage systems and media) and software (Operating System, Application, and File Format) technologies are critical to the long-term preservation of and access to digitally-stored data. What technologies have we chosen for our model, and why?

Many current and known, near-term technologies have been and continue to be discussed for archiving image content. Each has advantages and disadvantages.

An agnostic model

The model presented here does not rely on any particular technology. Rather, we are presenting a model to archive content digitally for very long periods of time – centuries, millennia. Choices made to implement this model will have a direct impact on the ability of that implementation to achieve the goals of the archive, to keep content in a pristine, accessible, and renewable manner for archival durations.

These implementation choices will reflect business decisions made within the industry, enabled by technological capabilities either current or anticipated – or both. Implementations of a Digital Content Archive will not be defined successfully by the IT industry; they must come from within.

The intent in developing this model simply is to show that digital archiving can be done and that it can be done cost-effectively using Commercial Off The Shelf (COTS) technology.

For this model we chose an Enterprise class tape library and tape media, but that choice was for modeling purposes only. The selection of hardware, software, file formats, etc., is outside the scope of this model. Choices between many technologies can be made; each may have its own pitfalls. This model however, is

agnostic as to implementation. Whether disk or tape or holographic storage, whether Open Source or proprietary software, implementation choices are immaterial to the subject of this paper – archiving movies digitally.

Hardware

As our model is built on tape, let's review the major current and near-term storage technologies. The ability to archive movies digitally is not dependent on the technology chosen, however. We have made technology choices for the purpose of arriving at a cost comparison, without which we simply have an unsolvable debate, rather than the beginnings of a business discussion.

Computer disk.

Disks have the advantage of very fast read and write access. They also have a high areal density, the amount of information, or bits that can be stored per a given area, normally expressed as bits / mm^2. A disadvantage of traditional disks is that they require electricity and cooling while running. Recent advances in a disk implementation called "MAID", or Massive Array of Idle Disks, address this to some extent in that MAID disks power-up and spin-up only when data is requested from them, returning to an idle state when not in use. Another disadvantage of disk storage is that it is more expensive than other media, such as computer tape. Computer disk has a long useful lifetime, arguably measured in decades, though its practical life is shorter due to the economics of increased density and error performance, and decreased electricity requirements due to advances in motors and drive electronics.

Computer Tape

Tape is the current medium of choice in nearly all industries and governmental organizations for the long-term storage and archive of digital data. Though slower in read-write access, tape requires neither electricity nor cooling when not in use and is less expensive than disk per data quantity of data stored in a given volume. In an archive environment in which media may be moved around to replicate deteriorated copies, or to return data to production at another location, tape is more robust in transport than disk. (Files of the size of movies will not be transmitted efficiently anytime soon.)

Holographic

Holographic storage, thought to be a near-term future medium for the archival storage of digital information, can store data inexpensively and densely. In current implementations its read/write rates are slow but it is likely that access rates will improve over time as this technology matures. It may be able to store data longer than either of the above media types, but that currently is unknown. Estimates of 50 or more years of viable data retention have been presented by vendors of this technology. (The viability of any technology over that time span is questionable and is dealt with below.)

Software

Operating System

Although several Operating Systems currently exist, the logical choice for a Digital Content Archive is an Open Source Operating System, of which there are two primary choices today: Linux and Solaris. Both are forms of UNIX, developed by Bell Labs in the 1960s and distributed for several decades for a nominal fee to academic institutions. Linux has many flavors and is supported by nearly all system manufacturers. Solaris is an implementation of Sun Microsystems, which provides it in a free-to-download and use model. In both implementations support is fee-based.

Open Source, whether in the Operating System, the Hierarchical Storage Manager (below) the application writing the data, or the actual file format of the archived data, should be strongly considered when choosing an archive system. An Open Source software environment likely will provide long-term viability for software distributed under it.

Hierarchical Storage System

The Hierarchical Storage System, or HSM, is a key software element of any digital archive. The ability to automate making copies of files, audit files for bit errors, reject bad files and make new copies based on the results of those audits, and to read-in an older file format and write-out a new file format (thus migrating the format and application information required to ensure archival integrity of the stored content) is critical. These functions arguably can be performed by a human, but it is certain that a piece of media will be missed, a migration not done, an error due to lack of training, sleeplessness or just forgetfulness, will occur.

These problems should not be allowed to have an impact upon irreplaceable archives. They easily can be avoided through the implementation of a simple, policy-driven tool, the Hierarchical Storage Manager. Additionally, the implementation and use of an HSM can lower the number of professional librarians required to maintain the archive, and hence decrease archive personnel costs.

For the same reasons as the Operating System, the HSM should be an Open Source application to ensure readability and support over archival durations.

Application Files and Formats

The format of the file written to the Digital Content Archive also should be open. The longest-lived open file format at this time is TAR, a UNIX file format. TAR and TARBALL files are readable even without the application which wrote them, a necessity in the ever-changing world of digital technology.

Regardless of the technology chosen for an archive, it will change over time. That fact is both the boon and the bane of technologists – things get smaller, faster, cheaper and better, but constant change is the price paid. It is doubtful that digital technology ever will cease maturing. Can we store a bit in the orbital path of an electron? Only one? Is this format or application or Operating System better than that one? In order to deal with this change, a migration plan must be in effect and followed.

There simply is no technology in the digital world providing for a static archive model; no storage or computing media that will remain constant enough for more than a decade; no application or format that will remain unchanged for even that long. A 50-year storage device or media? What physical reader will be able to read it, what application able to understand it, what chip still will have the instruction libraries to run the application through which the data was created and stored?

Our model

The choices we made in designing our Digital Content Archive model simply represent one implementation of such a model. These choices are necessary in order to discuss throughput, storage capacity, growth and to create cost models. They are not necessary to the below discussion of technology used to archive digital information over archival durations.

Migration of all of the model's technology, hardware and software, is built-in to the model.

For the storage portion of our model we chose computer tape for the archive, and an enterprise-class tape library. We chose this due to reasons of cost, maturity, and heating and cooling requirements required per petabyte stored and current automation.

Some current enterprise-class tape libraries have another attractive feature for a Digital Content Archive: they can store various types of tape drives. This makes them a logical choice through which the choice of media technologies can be fairly agnostic, leading to a more open environment, and the ability to change media types and vendors as necessary. As media technologies continue to evolve, it is likely the library will keep up. Putting a holographic cartridge, for example, into a tape-library form factor will enable that storage format to take advantage of the very large and quickly-increasing implementations of media robotics and automation as driven by the IT world at large.

The Object

Whatever is being archived and however it is archived, it is some *specific* object which is being archived. What *is* that object? Are there *multiple* objects? Are those multiple objects stored *differently* from one another? *How* are they stored? If these objects are stored differently, *how* and *why*? Is *everything* that is "archived" really *archived*? Does each of these objects have the same re-use pattern?

The Film Archive Object

In the current film archive model and process, four (visual) objects are stored.

1. The first is the conformed *Original Camera Negative*, or OCN. This may be about 6 - 2,000-ft cans of film. This is *stored* on color film.

2. The second is a *Color Match* print of the conformed negative, or another 6 2000' cans of film. This also is *stored* on color film.

3. The third object is the entire OCN, *stored* in color on film

4. The final object is the *Color Separations* of the conformed OCN, and is 3X the conformed negative, or about 18 cans of film. It is *archived* as black & white film through the process of color separations (simply: recording separately to B&W stock each of the colors Y, C and M through filters, recording the primaries onto B&W film with its far longer archive life; the result is three versions of the confirmed movie, one in each color, filtered onto B&W.)

In a deep archive Color Separations are managed in controlled environments by professional film archivists who do an excellent job with this irreplaceable material. The film is in archival cans with molecular sieves, and is kept in cool, dry environments to reduce or eliminate airborne pathogens and to slow chemical deterioration.

That leaves normally about 85-90% of the total footage shot *stored* as color negative or positive in long-term storage lacking the same protections of an *archive*. It is not protected by color separations and deteriorates more rapidly and unpredictably.

That we don't do color separations on the entire footage shot, the entire unconformed OCN, is evidence that we recognize already that we need not keep every frame of every take of every movie forever. By doing so we are accepting that degradation will occur with the unreleased portion of the movie – *and stating that we are OK with it*. Any degradation that does occur will be unpredictable, e.g. we will lose some of the image to degradation, but we don't know exactly what, how soon or how much. "Leaving aside the well-known problems of nitrate film we are faced with the stability problems of our polymer-based supports. All polymers are subject to decay." [6]

Managing the degradation of film and images recorded thereon is what the current science and art of film archiving is all about. But it is *managing* the degradation, not *eliminating* it.

Whether a color print or negative in cold storage, or B&W separations, film decays. "It is also clear that it is impossible to expect chemical inertness in the materials and that all will degrade. We can only hope to retard the degradation and preserve what we have as long as possible." [7]

A Digital Content Archive model with migration of the archived story to new media on a scheduled and reasonable basis will provide an environment in which the archiving of the *story* is not subject to the longevity of the *medium* on which the story is told. To use a relatively new Internet term, the archiving of the story is *disintermediated* from its archive medium.

This is as it should be – a story is not a physical thing; why should archiving a story be reliant on the characteristics of any physical medium?

The Digital Archive Object

Debate continues within the various standards and professional bodies within the American moviemaking industry (AMPAS, SMPTE, ASC, others) on the form and format of what we will call the "Archive Object,", e.g. the content to be digitally archived from a movie.

It is not the intent of the authors to define the format, the size, the form, or the amount of "footage" from the Original Camera Negative (OCN), or its digital equivalent, that will be archived. Those all are business decisions and are the purview of that industry.

Information Technology can *inform* those decisions, and models easily can be defined showing the costs and performance of various implementations. What ultimately is placed in digital archives, however, is a business decision that must be made by the industry whose content is being archived, not by the vendors of digital archive technology.

In the current *digital* feature production world, *what* is the Archive Object? That is not yet known; it is what is under sometimes fiery debate at the moment.

To understand the problem, let's look at a recent release, Zodiac (David Finch; Paramount-Warner Bros. joint production, 2007).

Zodiac was captured at a resolution of 2K. At that resolution its 18.2M frames consumed approximately 100TB of data. In order to be released to film-based theaters, the final version of the movie, the equivalent of the conformed negative (called the Digital Intermediate or DI) was output through a film recorder and recorded to film. (For a digital theater, the DI is used to create the Digital Cinema Distribution Master, or the DCDM.)

Is this a 2K Archive Object, digitally archived, a faithful and exact recreation of its capture? Do we archive all 18.2M frames, all 100TB? Do we archive it at the capture resolution of 2K? Or do we up-rez the 2K Digital Intermediate to a 4K resolution, creating an Archive Object of about one petabyte (1PB), 90% of which is made-up mathematically?

Or do we film-out and scan back at 4K all 18.2M frames, creating that way the 1PB Archive Object?

Or do we scan back from film only the released movie version at 4K? A 4K version of a 158-minute movie is about 12TB of data.

Or do we archive it on film? If we archive it on film, do we archive the entire OCN-equivalent 18.2M frames via film-out, approximately 1.5M feet of film, *and* a color film-out of the DI, 227,000 frames or 14,220 feet

of film, *and* a color match print – another 14,220 feet of film, *and* do color separations on the film-out of the DI, another 42,660 feet of film, a total of 1.6M feet, or 786 2000-ft cans of film?

Do we do this at the capture resolution of 2K, or do we up-rez all of the 18.2M frames algorithmically to 4K and then film-out and archive *that*?

If for any reason we lose the DI or the seps, or want to add footage for a new release, what is the original we go back to? A film copy of the digital capture? The digital files themselves? At what resolution – their captured 2K or the up-rez'd 4K, or the 4K scanned from the film recorded at 2K or 4K?

To try to wade through this, let's look at what is done today and figure out the *digital equivalent* of each of the objects archived today, using the same list of four objects as above, and the same movie, Zodiac:

1. The conformed *Original Camera Negative*, or OCN. The digital equivalent is called the DI. But this was a 2K capture, and we will be able to up-rez as well in the future as we can now, perhaps better, so do we archive the 2K DI at approximately 2TB? Or a 4K DI at 12TB?

2. The *Color Match* print is not necessary because mathematically this would be exactly the same as #1 (except a negative rather than a positive, which is created mathematically).

3. The *Color Separations* of the conformed OCN is not needed for the same reasons as #2. The color seps are three copies for archive of the conformed OCN. In a digital world we would keep instead multiple copies of the DI. (In our model we propose to keep four copies.)

4. The entire OCN equivalent, scanned to data (we still aren't solving the 2K – 4K issue however). If we scan it at 2K, we have another 100TB; if we scan it at 4K, another 1PB.

The only problematic item in the list is the last, and strictly for reasons of size. In any instance, regardless of the size of the object, a digital archive, using established data archiving Best Practices, will return at any time in the future *exactly* what it was given, bit-for-bit.

So let me repeat a statement from above…

> *That we don't do color separations on the entire footage shot is evidence that we recognize already that we need not keep every frame of every take of every movie for archival durations… and that we are OK with that*

…an d translate it into a digital world.

By *not* performing color seps on the entire OCN, we are currently stating that degradation is acceptable over the long term for that OCN, but that we want to keep the release print (and whatever other elements on which we do color seps, such as a Director's Cut), for archival durations.

Why would we want to keep a high-resolution digital-equivalent-OCN forever? Will any version other than the released one (or Director's Cut) *ever* be repurposed theatrically? The original version may indeed be re-released theatrically, and so must be kept at it highest resolution (which still doesn't answer the question: Should a movie captured at 2K be archived at 4K?), but will any other version *ever* be repurposed at a resolution higher than HD?

This line of argument is anathema to preservationists, but may be worth exploring. Let's look at compression.

The IT industry can do lossless compression at about 2:1. That would give us, in a 4K OCN, an archive object for an up-rez'd, future-proofed Zodiac digital OCN of about 500TB, or an archive object of the original 2K capture of about 50TB.

Since the highest-resolution repurposing device likely will be HD, however, and since the highest bit-rate for HD is Blu-ray®, then the highest bit-rate for repurposing is on the order of 20Mbps lossy compression. So a lossy compression of the entire 4K OCN at, say 100Mbps, or 5X Blu-ray, would enable "future-proof" digital repurposing of the entire OCN equivalent at 5X Blu-ray's capacity. That 100Mbps compression would yield a file size of 12TB

What about that "future-proofing"? What about future displays of a resolution higher than HD? We can't store at 100Mbps when we may need the uncompressed OCN at an uncompressed 4K to meet future resolution requirements, right?

Several arguments militate against this.

1. The example in use, Zodiac, was captured at 2K. Making it 4K to "future-proof" it using math creates something not captured or posted, but rather something entirely new and subject to whatever faults any up-rezing program may induce. This is not a desirable effort in an archive.
2. The authors of this paper are not experts in the human visual system, but from conversation with industry experts, the ability of the human visual system to resolve images ends somewhere between 3K and 4K, digital equivalent. (Yes, the human visual system is analog and not digital, and no exact comparison can be made.)
3. Our ability to resolve images depends on screen size, however, which is why the conformed OCN, or the DI, is archived as color separations, and logically should be archived digitally in a 4K lossless format. It is doubtful that anything other than the original release, or at most a Director's Cut, *ever* will make it to the big screen once the movie has passed its theatrical window. Given that there is no on-screen repurposing of more than the Director's Cut conformed negative, or any other repurposing done now in any non-digital (or uncompressed) format, and that consumer repurposing format requirements are getting smaller, not larger (cell phones, not 103" TVs at home – though even those are only HD resolution), the chances of future repurposing at more than 5X HDTV may be vanishingly small.
4. Another argument is financial. The concept that – at *any* time in the future – consumers in numbers sufficient to cause a market transition to resolutions higher than HD, that they (or broadcasters) will be persuaded to exchange HD TVs for higher resolution, is problematic. The money now being spent to convert broadcast and televisions to HD will not be spent again; the return for a perhaps imperceptible increase in resolution, an increase not visible to most humans, is so small as not to be able to drive the consumer, or the CE manufacturers or broadcasters in that direction.

But even so, why would we accept a compressed version of an irreplaceable object?

1. It makes perfect sense that the object telling the story – the DI or the scan of the conformed negative (or Directors' Cut) – will be uncompressed at 4K. *That* is the cultural artifact; *that* is the revenue vehicle. *That* is what color separations are done on now. (Until one argues that up-rezing may be better in the future than now and that one should archive at the 2K resolution shot. Migration to 4K capture will moot this issue.)

2. *Not* doing color separations now on the entire OCN means we *accept* that it will degrade, even knowing it will degrade unpredictably. (If we thought that the OCN wouldn't degrade, even when refrigerated and dehumidified, with molecular sieves, we wouldn't be doing color separations.)

In a high-bit-rate lossy compression, on the other hand, we can determine *exactly* the "deterioration" or image loss we are willing to accept, make an object with *exactly* those qualities, and then keep it pristinely – forever.

In order to arrive at a cost and performance model, however, and for the purposes of modeling alone, let's review the major options being considered today, and their object sizes. We'll continue to use Zodiac for these calculations.

(In all cases of digital archiving, metadata, EDL, CDLs, color spaces, etc., must be archived in addition to the actual content. The size of this information is so small relative to the movie size, however, that it is swallowed up by the overall object size of the movie to be archived, so is ignored in the calculations below. It will, however, be archived and this statement is not meant to dismiss this additional data or its critical importance.)

1. Store the entire OCN equivalent at the resolution at which it was captured and posted, uncompressed, and the DI at 4K (112TB)
2. Store the entire OCN equivalent at 4K, regardless of capture and post resolution, uncompressed, and the DI at 4K (1PB)
3. Store the DI at 4K uncompressed, plus the entire OCN at capture resolution, 100 Mbps lossy compression (Zodiac, 2K = 12TB) (24TB)

Should 4K capture become the norm, and should cameras be left on between takes, as some suggest, the numbers above will increase – drastically. In addition to file sizes, this is problematic for another reason: The cost of such an archive – film or digital – is completely unpredictable, but certainly very high.

One industry executive has stated anecdotally that the size of the Archive Object will be dealt with as a business decision: "You get 400TB, put in whatever you want." Though it sounds flip (and the 400TB likely was not scientifically derived), this comment is driven by business logic which, as Ms. McCargar points out, increasingly is driving archive decisions.

Our Archive Model

The Heart of the Debate -- Ensuring the Bits are still the Bits

Hollywood is accustomed to archive durations for their movies in excess of 100 years, far longer than archives required by federal and state authorities for various other purposes, such as tax audits. One can re-build bank records from general ledgers, and retype or re-scan printed material from paper (which has millennia-long archival periods when properly stored). One cannot, however, re-record Louis Armstrong or Buddy Holly, or re-shoot Henry Fonda, Jimmy Stewart, Audrey Hepburn or Greta Garbo.

Once the hardware and software technologies and implementations have been sorted through, once the Archive Object has been defined, how do we, using digital technology, store, provide access to, and provide infinite renewability of that irreplaceable Archive Object?

Ensuring Data Integrity Algorithmically

Algorithms and procedures exist to ensure bits stored digitally remain the same bits over extended periods of time. As our model uses a tape media implementation, we'll start with Error Correction Codes, or ECCs.

When writing bits to a computer tape, tape drive hardware uses an ECC(8), methodology to encode the data so that if fingerprints get on the tape or the emulsion flakes off, or even if holes are punched in it, the drive will detect this and reconstruct the data from redundant bits recorded elsewhere on the tape.

When the tape drive actually writes the data to tape, in an archive application it must be set to perform read-after-write, in which it reads back that data immediately to ensure that what it was told to write is what it really did write. The drive software writing the file uses the ECC to accomplish this. Though this process slows the overall data throughput, not providing error-correcting capability for an archived object is a mistake.

To have recourse in the event that data on a particular tape cannot be read in the future, it should be ensured that two copies of each Archive Object are made and stored in the Digital Content Archive. For further reliance, another library archive, known in IT parlance as a "Disaster Recovery" or "DR" site, should be maintained in a different geographical location, and that another two copies of each Archive Object are kept in that remote library. Four copies of each Archive Object thus are archived.

Independent studies of image archive needs and technologies exist. A recent one performed by SAIC [11] for the US Government, is *Offline Archive Media Trade Study*, [12] for the USGS (United States Geological Survey) at EROS (Earth Resources Observation and Science), and published in FY01 and revised in FY03, FY04 and FY06 (October, 2006). In it the authors state the following:

> When two or more copies of a dataset exist, and one is already on an enterprise technology, use of an enterprise solution for the second copy is not warranted. [13]

The problem with this statement for the movie industry is one of the migrations of very large objects.

One certainly can choose to provide only one automated tape library, storing a second (or third) copy on a shelf elsewhere. Because it is desired to keep these Archive Objects pristine for centuries, however, they must be migrated. If the thousands of media cartridges at the Disaster Recovery location are not in a library, an enormous amount of expensive and error-prone human labor is required for each migration. In our view it is better by far to automate both locations, and schedule the Hierarchical Storage Manager to perform these audits automatically, than to rely on a human to accurately perform each migration. Hence we have two locations with identical tape libraries, computer and disk front-end, etc.

Industry Experiences and Current Technology

As the first version of this paper was presented and discussed, many knowledgeable people in the industry pointed out that the error rates quoted by the IT vendors, and quoted from them in this paper, were not what they had experienced in their companies. Before we go into the Bit Error Rates (BER) of the technologies we modeled, these real-world experiences need to be addressed.

In order to do this, let's turn again to the Media Trade Study performed by SAIC, and most recently updated in October, 2006.

In this Study the elements defined as most contributory to data reliability were: [14]

1. The number of archival copies
2. The storage location and environment
3. The composition of the media

4. Tape handling within the media
5. Error handling

In our model we addressed these as follows:

1. Four archival copies
2. Two locations, environmentally sound for tape library performance
3. Advanced Metal Particle (AMP), with a verified 10^{-19} uncorrected bit error rate [15]
4. A drive type in which the recording surface of the tape never is touched by the drive, and which is recorded longitudinally rather than via the method of helical scan ("The [archive] technology must not be hampered by a poor reliability or performance history. Helical scan technologies such as 8mm, 4mm, DAT and D3 have proven unreliable in the past." [16] "Helical scan technologies have proven unreliable in the past due to complex drive path, high and constant head contact, poor transfer rates, and extremely poor start/stop/repositioning times. Ancestor 8mm technology was prone to destroying tapes due to mis-queues in controlling the complex tape path. Time will tell whether this heritage has been overcome.") [17]

   (Among technologies "dismissed" from this study from further analysis or consideration," [18] were: CD-ROM, DLT 8000, QIC, Mammoth, Erasable Optical, Sun/STK 9840C, Exabyte VXA320 "(and similar 4mm/8mm helical scan technologies)" [19], DVD, HD-DVD, Blu-Ray, "Newer optical technologies". [20])
5. The use of ECC error correction technology

(Note: The above-referenced trade study is a comprehensive review, publicly available, of the major archive technologies. We would recommend referring to it when investigating this topic.)

Addressing the industry comments with regard to experiences of only months or one or two years of storage before a tape became unreadable is problematic. In fact, if these experiences were widespread across the IT world, tape drives would not be used for any storage whatsoever, and tape drive and media manufacturers would be out of business.

That tape systems store the vast amount of stored data in the world today argues that perhaps something other than tape itself, and digital technology generally may be the reason for these failures.

iMation© published in August of 2003 a paper, "The Storage Lifetime of Removable Media (Mary Chester, iMation, August, 2003). The paper includes a short summary of Best Practices regarding the handling of tape cartridges:

> **Handling and Care of Tape Cartridges**
>
> Together with the hardware manufacturers, media manufacturers recommend the conditions under which tape cartridges should be handled, transported, stored, and used, in order to insure that the product will perform to standards over this 15 to 30 year time period. Data can be affected by environmental factors such as debris, high temperature or humidity, drastic temperature or humidity changes, and stray magnetic field sources, as well as improper handling of the cartridges by either operations personnel or by the hardware. If not properly handled, high-capacity tape cartridges are susceptible to damage due to the increased linear density, increased track density and subsequent positioning of the data and servo tracks closer to the edges of the tape.
>
> Some basic rules for handling tape cartridges include:

- Stack or carry no more than six cartridges at a time to minimize the risk of dropping the stack.
- Do not place cartridges that are dirty or damaged in a drive.
- Use the finger grips, if present, to carry a single cartridge.
- To prevent tape damage, do not remove leader blocks or open drive doors.
- Do not touch tape surfaces, as residue from a fingerprint can create greater head-to-tape separation and result in loss of signal (data).
- Respond to drive messages for cleaning as directed, and only use cleaning cartridges recommended by the hardware provider.
- Assure that drives are maintained and serviced per the hardware manufacturer's specifications.

Additional topics, such as dropped cartridges, transportation and storage of cartridges, and operating environments also are covered in this brief paper.

Referring again to the SAIC/USGS Trade Study, many factors contributory to errors exist: Media composition [21], the complexity of the tape path [22], the professional or consumer grade of the media and drive technology [23] , and many more. Simply put, tape is not tape is not tape, and drives ranging in price from $500 to $25,000+ certainly are not identical.

IT-based processes also come into play. "The project managers from the "cutting edge" projects emphasized the importance of considering best practices for archiving at all stages of the information management life cycle." [24] With these Best Practices followed, those industries archiving digital images, such as Oil & Gas, Geospatial, Geologic, and others, achieve published, verified Bit Error Rates. [25]

Bit Error Rates

To ensure that the movies in the archive are as close to "perfect" as possible, two things must be accomplished with the Archive Object: Bit failures must be *detected*, and then they must be *corrected.*

While *undetected* BERs are higher than *uncorrected* BERs, and while many excellent mechanisms exist to *detect* bit error loss, it's not enough in a Digital Content Archive to *detect* errors – it must be possible to *correct* them as well.

Industry verified *uncorrected* BERs range from 10^{-19} for Sun's T10000 Tape [26] , to IBM and Hewlett Packard's LTO-3, having a verified *uncorrected* BER of 10^{-17}. (The new HP LTO4 lists an unverified 10^{-17} *uncorrected* BER. [27])

This means that a 10^{-19} (or 10^{-17}) chance exists of getting an error that the tape drive's ECC cannot repair per movie archived. These nines are additive. With four copies of the Archive Object, the chance of an *uncorrectable* bit error across all four copies of the Archive Object in the archive is 10^{-76} (10^{-19} to the 4th power). This seems to be reasonable protection for an archive: that's 76 nines of protection against all four of these movies suffering a single bit error that cannot be corrected.

The Time Factor

Like film, data tape tends to degrade over time, and a century is a long time. Because of the irrepressible dynamic interactions of electromagnetic signals and the physical tape medium, what Mssrs. Hunt and Hummel referred to at JTS2004 as, "the fugitive and complex nature of electronic systems," degradation of the on-tape signal (bits) can occur more quickly than images on film. A media cartridge – *any* media cartridge - in and of itself is not conducive to a century archive.

Additionally, the form factors that digital media take evolve at a rate requiring one to move digital files "forward" to new media and form factors, a need that must be met to ensure long-term ability to have systems capable of reading the Archive Object. These physical problems and technological realities create the need to re-write each digitally-stored Archive Object on a periodic basis.

The current top-of-the-line tape cartridges, for example Sun's T10000 and IBM's 3590 Extended High-Performance Cartridge Tape[28], have published archive lives of 30 years. Additionally, the National Media Lab uses a duration of 10-30 years as the usable life of magnetic tape [29].

> My NML colleagues and I agree with the key point made ...t hat the technological obsolescence of digital recording systems is a challenge for those individuals tasked with preserving digital archives. Digital archives should be transcribed every 10 to 20 years to ensure that they will not become technologically obsolete... Life expectancy estimates of 10 to 30 years for magnetic tapes are common.

In a digital world, however, *nothing* has a 30-year *useful* life. Quoting again from the above-referenced Trade Study, [30]

> The technology must use a media that can remain readable for at least 10 years in a controlled environment. The lifetime of 10 years was selected since it is the longest that a media technology would conceivably be used before space and transfer rate concerns would dictate a move to a new technology.

(Taking a closer look at the economics of library space, the Trade Study reports that the USGS migrated more than 50,000 tapes over a period of 5.5 months, and, "This migration freed up enough library shelving [slots] to ensure that the library should never need to be expanded, and may in fact be reduced in size." [31])

In our model we propose to rewrite each Archive Object to new tape cartridges once every 5 years and whenever cartridge density changes significantly. (We propose a full tape audit every six months, however, to validate data integrity.)

At the same time, one's 'mileage may vary,' and the stated use by some of 5-7 years as a re-recording duration differs from the 5-year life we have modeled. Going back to the NML study and response,

> Experience indicates that physical lifetimes for digital magnetic tape are at least 10 to 20 years, a value commensurate with the practical life of the digital recording technology. One government agency responsible for maintaining meteorological data archives recently transcribed approximately 20,000 ten-year-old 3480 tape cartridges, of which only two cartridges had unrecoverable errors. Properly cared for reel-to-reel, 9-track computer tapes recorded in the 1970's can still be played back in the 90's, even though the 9-track format became obsolescent in the 80's. The NML has investigated the stability of several forms of digital storage media over the last six years. Life expectancies for magnetic media can be estimated by modeling the deterioration of tape properties induced experimentally in accelerated aging environments. Life expectancy estimates of 10 to 30 years for magnetic tapes are common. Given the fact that digital recording technologies can be supplanted by a newer format every 5 to 10 years, the bigger problem facing archivists is the lifetime of the technology, not the lifetime of the medium.

By rewriting the Archive Object periodically, we are accomplishing two major tasks: we are ensuring the currency of the tape cartridge and drive (i.e. we never will have a 100-yr-old cartridge and go looking for a machine in which to read it, nor a 100-yr-old tape drive looking for a library) and, as the tape software changes and upgrades over time, we are keeping current with that software as well (ensuring we never

have a 100-yr-old tape file format without any tape software to read it). (The application software that determines the *file* format is a different issue and is dealt with below.)

With a 5-year rewrite period, each recorded Archive Object will need to be rewritten 20 times during the course of a century archive. What will be the Bit Error Rate at the end of 20 generations of rewrite, and will that result in a usable movie capable of meeting a revenue-creation opportunity?

Here's the interesting part:

> *For this application, it doesn't matter how many times the data is accessed.*

Here's why:

The probability that the ECC's undetected BER will fail to correct detected damage during any given access is 10^{-19}. The probability that it will fail one more times during N accesses is 1 minus the probability that it will succeed N times in a row, i.e. $1 - (1 - 10^{-19})N$. For N less than 10^{19}, this is well approximated by $N * 10^{-19}$.

(For the sake of an exaggerated example, let's say a movie is accessed one million times. For an Archive Object the chance of misreading the movie (e.g. having one uncorrectable bit error) on one attempt is 10^{-19}, so the chance of misreading on any of the one million accesses is $10^{6} * 10^{-19} = 10^{-13}$.

The point being: it reasonably can be assumed, for the purposes of this application, that the ability to detect errors in transcription is perfect.

The Software to Archive the Content

But the software performing the moves, reads, copies, and migration *does* matter. Long-term storage of static data is the reason archive management software exists.

Hierarchical Storage Management (HSM) software provides the capabilities necessary to a digital content archive: policy-driven data movement, scheduled copy/migration, automatic copy creation, the ability to create and manage file subsets, etc.

Other strategies

While using ECC alone can provide significant levels of protection, other strategies exist for preserving high-value content over extended periods.

As described in this paper, the redundancy afforded by having four physical tape copies of the data is used only in an all-or-nothing manner: either a copy is good or it isn't. In fact, strategies exist involving majority voting (on each bit, if necessary) among three or four copies that can greatly improve the reliability.

Moreover, there are esoteric mathematical methods that could be applied to further increase reliability [32]. One could, for example, construct from a 10TB Archive Object 40 files of 1TB each with the property that the original file can be reconstructed from *any* 10 of the 40 files. This means that one is able to reconstruct the original file if at least 10 of the 40 files were uncorrupted – and each of those 1TB files would be protected by the ECC. Such strategies are computationally intensive, but in an archive environment, i.e. one not driven by high transaction rates, the computational costs maybe acceptable.

Weaknesses

The biggest weaknesses in any archiving scheme are external agents: human error, sabotage, network-based attacks, and natural disasters. Therefore our model suggests the following:

(1) Find ways to create institutional memory of the archive's existence and of the need to refresh it periodically. This seems obvious, but history suggests it's the weakest link.

(2) Have two or more sites, geographically far apart, maintained by different people.

(3) Have absolutely no network connectivity. Zero network connectivity would be desirable in order to preclude, with the exception of the Archivist, any type of electronic tampering or theft, internal or external. Lack of connectivity would be balanced against the costs of physical transfer of these tapes and the security involved therein, but films now are transferred via insured physical transfer. (The intent in proposing a Digital Content Archive isn't to break working models. It is rather to enhance those models. In fact, however, while encrypting an archive would not be a wise decision, i.e. the archive files always should be in the clear (losing the key or the algorithm or the encryption application could render the entire archive useless), a tape recalled from archive certainly could be encrypted for transport and then decrypted once it reached its destination. Hardware encryption on the tape drives is a reality now, adding very little performance cost to secure the content during transit.)

(4) Each time a new copy of a movie is generated, send the old copy to another location. It costs almost nothing; doing so verifies the movie still is intact; and it may remain intact for quite a bit longer.

Retrieval from Archive

If, based on the above, it is decided that movies can be archived digitally, what about the recall and reuse of those assets?

Recalling a digitally-stored archive copy will cost a small number of hours (which will decrease as tape density increases, throughput being related to data density on the tape), and can be delivered within a day (from the local archive) to a disk drive subsystem available to the person requesting the content.

(The latency involved in local delivery is determined by the throughput of the tape and disk subsystems. With current-year (2007) Enterprise Library and drive throughputs of approximately 120 Mbytes per second, it may take as long as 1389 minutes, or 23 hours, to read a 10 TB DI file.

(Multi-threading can decrease this time, but not to insignificance. It may be tangential to the main argument of the paper, but decreasing fetch latency by a factor of four or even two, might be of great value to consumers of the archived data. As some HSM systems can perform multi-threaded read-back, if this is desired it should be taken into account when designing the Archive.)

A copy of the movie from the remote archive never should be requested by a user. The remote archive exists to ensure the currency of the local archive, which should be used to respond to all requests.

In a digital archive, no time is spent on temperature and humidity standardization, finding a single physical asset in a large salt mine holding tens of thousands of identical film cans, insuring and shipping it, and getting a usable image to the desk of the requestor. (As the digital image is pristine, nor is time lost or money spent on restoration.)

Reuse of Image

Nor is any time or money spent doing a new scan of the color separations, the required first step in repurposing from the archived movie. The delivered asset already is digital.

But is it in the digital form required? What about the software used to write-out the Archive Object to begin with? Is that software still on the market? Can it still read the down-level file created years ago? Has the company which marketed the file format application software gone out of business or been acquired and the software no longer available? Is a computer on which it runs still available?

Fortunately each of these can be dealt with within the normal functioning of the technology world.

As a new file format displaces a popular format, backward compatibility is normal. In the few instances in which this has not been the case resourceful people have written and marketed backward-compatibility tools. Or the software in question has become so standard that similar, competing applications have created import tools. One major studio for a decade has been escrowing the source code for the software used to write their digital movie files as further assurance of availability.

Assuming the availability of the operating system under which this software runs, and a computer on which the operating system can be executed (see below), these and the above file format issues can be dealt with in the above-described generational refresh of the tape drive.

Above we proposed that the archived content is re-written every five years. Should the future of the legacy application software (or operating system or computer hardware running the operating system running the application) be tenuous at any time, the Archive Object can be ingested from the archive into a then-current application and system, and re-archived in that new file format. In fact, this process logically would be tracked and scheduled to ensure that as the industry migrates to new file formats the appropriate steps are taken to archive the content in that new format. This is feasible because the underlying data format of the Archive Object remains digitally fixed; only the storage medium changes.

Costs

If we assume from the above that one concurs that a movie reliably can be archived digitally for the required durations, what are the costs? After all, the scenario above posits that many generations of tape cartridges and tape drives will be required, a tape library to hold the drives, and computing power and front-end disk storage to run and manage the process. The disk and compute front-ends for the tape systems will reach end-of-life approximately every five years and will need to be replaced or upgraded, the tape drives every five years (or when density changes significantly and the change is economical), and the tape libraries holding those drives and cartridges will require replacement every twenty years or so. What are the impacts of these recurring costs?

Can Digital Costs Compete with Film?

A comprehensive answer to this would require consensus on the Archive Object. Is it 10TB? 100TB? 500TB? 1PB? More? Or will it be the above anecdotal business decision, “You get 400TB; put in whatever you want.”

For the purposes of this investigation, we chose a 100TB Archive Object. The accuracy of that choice is unknowable, but it provides what seems to be a reasonable place from which to start the model and discussion.

Film-based 100-year Archive Cost

Let’s look again at the four items archived and stored in the current film-based archive process, extending the current approximate costs to store each for 100 years (costs are from a large post-production company and are approximate):

1. The first is the conformed *Original Camera Negative*, or OCN. This may be about 6 - 2,000-ft cans of film.
    a. 12,000 feet of color film, processed = $7,800
    b. $1.00 per can per month
    c. 6 cans X 1200 months X $1.00/can/mo = $7,200

   d. Conformed OCN film-out and store: $15,000

2. The second is a *Color Match* print of the conformed negative, or another 6 2,000-ft cans of film.

   a. 12,000 feet of color film, processed = $7,800

   b. $1.00 per can per month

   c. 6 cans X 1200 months X $1.00/can/mo = $7,200

   d. Color Match print film-out and store: $15,000

3. The third object is the entire OCN, with an approximate average of

   a. five pallets of camera negative in 1,000-ft cans, stored in "Dead storage"

   b. Cost of $700/yr = $70,000

4. The final object is the *Color Separations* of the conformed OCN, and is 3X the conformed negative, or about 18 cans of film. It is *archived* as black & white film through the process of color separations (simply: recording separately to B&W stock each of the colors Y, C and M through filters, recording the primaries onto B&W film with its far longer archive life; the result is three versions of the confirmed movie, one in each color, filtered onto B&W.)

   a. 36,000 feet of B&W film, processed = $8,000

   b. $1.00 per can per month

   c. 18 cans X 1200 months X $1.00/can/mo = $21,600

   d. Color separations film-out and store: $29,600

The total cost to archive a movie for 100 years using current film-based practices and costs is approximately: $130,000.

If we want to repurpose this movie only one time in the century, we must re-scan the color separations ($65,000). If any restoration work is required, that is an additional cost.

Adding in the cost to re-scan for repurposing one time in that century, we have a cost of $195,000 per movie archived and repurposed once per century.

If any restoration is required on the film, that is extra and normally bid on a per-frame basis and is in addition to the above cost.

Digital 100-year Archive Cost

Using the proposed Digital Content Archive model, a 100TB Archive Object, including migration, power, three Librarians/site (loaded annual salary of $130,000 per Librarian) the cost for a 2000-movie archive is approximately $73,000 per archived movie per century, based on a single-studio contribution of 20 movies

per year added over the century. (This cost assumes a Sun/STK T10000 drive; using LTO the cost would be approximately $66,000.)

The cost to prepare a movie for repurposing (in a digital archive simply the cost of the cartridges to make a new copy, using 1TB cartridges and a 10TB DI) is approximately $2,000.

For one movie, then, assuming an archive of 2,000 movies over a century, the cost is:

	Archive	Repurpose once	Per-Movie Total	2,000-Movie Total
Film	$130,000	$65,000	$195,000	$390,000,000
Digital	$73,000	$2,000	$75,000	$150,000,000

Other Digital Image Archives

Digital images are at the core of industries other than entertainment. These include, but are not limited to:

- Oil & Gas
- Geospatial Imaging
- National and University archives
- Intelligence and Defense

Oil & Gas has been archiving images digitally for decades. They have evolved an industry metadata model and archive format. New oil often is found by running new algorithms against digitally-archived images from old but very expensive seismic shots. Archiving digital images accurately for decades is critical to this industry.

Several major libraries are archiving their media assets digitally: The American Library of Congress, The Stanford University Libraries (in conjunction with Cambridge, the Biblioteca Alexandria and the French National Archive), the California University system (34 campuses), the British Library, and Dalhousie University and McGill University in Canada all are archiving media assets digitally and have migration strategies in place similar to those modeled here.

The assets being digitized for archive by these librarians and archivists are as irreplaceable and as culturally and economically valuable as the stories told through the movies. That their archivists have chosen to archive them digitally speaks volumes.

Conclusion

The consumer wants their content anywhere, on any device at any time. Creating a Digital Content Archive as described here will enable the content owner to provide the consumer with the content they desire in a format they want. The current analog archive model simply can not meet this need.

The studios have archived analog content to license, but it is stored nearly inaccessible to the market and consumer. This makes questionable at best the ability of the content owner to realize added revenue on their archived content. Continuing with this model may not be in the best interests of content owners.

The consumer wants content. As was noted by several presenters at IBC in 2006, content no longer is "King." The mantra now has become, "The Consumer is King." To meet the growing and varied means in which a consumer will demand content (from in-theater digital prints to Video-on-Demand to cell phones and likely more ways than are dreamed of today), only a digitally-managed file can be reformatted and reused with mathematical purity and accuracy.

The ease-of-access to digitally-formatted movies for recall and repurposing is far higher than access to film archives. The steps to repurpose digitally-stored content are reduced to one. The time and cost to recall a movie from the archive are far lower. All of these factors result in increased use of and revenue from that archived movie.

Using a Digital Content Archive to make available studio content quickly and efficiently can create an entirely new business model based on the speed of the market adoption and the velocity with which content will flow through this new market to consumers in theaters, in homes and on mobile devices.

The archive can become a part of the revenue flow of a studio, rather than simply a large and ever-increasing cost to maintain.

Technology has matured to the point at which this Digital Content Archive is feasible. The Library of Congress has embarked on a similar long-term digital preservation and archive project for all of their media assets: "Storing National Treasures" http://www.enterprisestorageforum.com/sans/features/article.php/3586066.

and: "Sun Rises at the Library of Congress" http://www.enterprisestorageforum.com/sans/features/article.php/3619646.

The time to begin serious efforts on testing and implementing studio digital archives is now.

Notes

1. Keynote address, Society for Imaging Science and Technology conference, May 21-24, 2007, Arlington, VA [RETURN]
2. "MPAA's VALENTI OFFERS SUPPORT FOR INDUCING INFRINGEMENT OF COPYRIGHTS ACT OF 2004" Statement by Jack Valenti, President and CEO, MPAA, June 23, 2004 [RETURN]
3. MPAA "2006 Box Office Rebounds", March 6, 2007 [RETURN]
4. In this paper the term "Archive" is used distinctively. It does not mean long-term storage alone, but encompasses both technology and processes to ensure that content stored is preserved for generations, and can be recalled efficiently so as to be available to its owners for additional repurposing revenue opportunities far into the future. [RETURN]
5. "The Color Conundrum," Douglas Bankston, ASC, *American Cinematographer*, January, 2005. [RETURN]
6. "That Smell in the Vaults: The Degradation of Polymers in AV Materials," W. Mark Ritchie, KINEMA, Spring 1995 http://www.kinema.uwaterloo.ca/ritch951.htm [RETURN]
7. ibid [RETURN]
8. http://en.wikipedia.org/wiki/Error-correcting_code [RETURN]
9. deleted
10. deleted
11. SAIC (Science Application International Corporation), San Diego, CA [RETURN]
12. Architecture and Technology Program September 2006 Offline Archive Media Trade Study, Todd Bodoh, Sr. Systems Engineer, et al, SAIC USGS/EROS, Sioux Falls, SD 57198, September 2006 [RETURN]
13. ibid, p 12. [RETURN]
14. ibid, pp 13-14 [RETURN]
15. ibid, Table 2-1, "Technology Comparison," p 11. [RETURN]
16. ibid, Section 1.4.5, p 9. [RETURN]
17. ibid, Section 2.1, p 19 [RETURN]
18. ibid, Section 1.5, pp 9-10 [RETURN]
19. ibid, Section 1.5.3, p 9 [RETURN]
20. ibid, Section 1.5.5, p [RETURN]
21. ibid, Section 1.3, p7 [RETURN]
22. ibid [RETURN]
23. ibid, Section 1.3, pp 7-8, and Section 1.5.3, p 9. [RETURN]

24. "Best Practices for Digital Archiving, An Information Life Cycle Approach," Gail. M. Hodge, Information International Associates, Inc., *D-Lib Magazine,* January 2000, Volume 6 Number 1, Section 4.0 [RETURN]
25. Architecture and Technology Program September 2006 Offline Archive Media Trade Study, ibid, Table 2.1 p 11, Table 3-1, p 21 [RETURN]
26. deleted
27. Architecture and Technology Program September 2006 Offline Archive Media Trade Study, ibid, Table 2.1 p 11, Table 3-1, p 21. [RETURN]
28. ftp://ftp.software.ibm.com/common/ssi/rep_sp/n/TSD00259USEN/TSD00259USEN.PDF [RETURN]
29. http://palimpsest.stanford.edu/bytopic/electronic-records/electronic-storage-media/bogart.html , quoting from a *Letter to the Editor* to *Scientific American*, by Dr John C. W. Van Bogart, Principal Investigator for the Magnetic Media Stability Program at the National Media Lab, in response to the article, "Ensuring the Longevity of Digital Documents (January, 1995, *Scientific American*) [RETURN]
30. Architecture and Technology Program September 2006 Offline Archive Media Trade Study, ibid, Section 1.4.4, p 9. [RETURN]
31. Architecture and Technology Program September 2006 Offline Archive Media Trade Study, ibid, Section 1.2, p 7. [RETURN]
32. See, for example, M. O. Rabin, "Efficient dispersal of information for security, load balancing, and fault tolerance," Journal of the ACM, 36(2):335–348, ACM, April 1989. [RETURN]
33. deleted

About the Authors

Dave Cavena is a Systems Engineer with Sun Microsystems, supporting the digital content space for studios and postproduction companies. Prior to coming to Sun he held positions as Project Director, Blu-ray Authoring System for Sony Pictures Entertainment, Program Manager for The Walt Disney Company's MovieBeam, and as IBM's Digital Cinema Executive Project Manager. Dave can be reached at: david.cavena@sun.com

Chris Wood, Director and Chief Technology Officer for Sun's Data Management Storage Practice, Client Services Organization is responsible for identifying and delivering the best solutions available that can address our customer's complex data management problems. He joined Sun Microsystems when his prior company, MaxStrat, was acquired by Sun early in 1999. Mr. Wood has held prior positions at IBM, Litton Industries and other computer-related firms. He can be reached at: chris.wood@sun.com

Jeff Bonwick, DE and CTO, Storage, Sun Microsystems. Jeff can be reached at: jeff.bonwick@sun.com

Guy Steele is a Sun Fellow with Sun Microsystems Laboratories, conducting research in programming languages, algorithms, and processor architecture. He is a well-known author or co-author of books about the programming languages C, Common Lisp, High Performance Fortran, and Java. He is also an ACM Fellow and a member of the US National Academy of Engineering. Prior to coming to Sun he was a Senior Scientist at Thinking Machines Corporation, a pioneering manufacturer of massively parallel supercomputers and of the DataVault, the first commercial RAID disk array. Guy can be reached at guy.steele@sun.com

Mike Selway is a Consulting Systems Engineer with Sun Microsystems supporting the crafting of data management tiered storage solutions around Sun Microsystems' high performance storage management file system, SAM-QFS. He has been a member of several data storage organizations integrating a wide variety of hardware and software technologies for creating film, audio, video, and post-processing market-specific data management solutions. Mike can be reached at michael.selway@sun.com

Acknowledgements

This paper was a collaborative effort involving many Sun Microsystems professionals. The authors would like to thank the following contributors for their assistance in reviewing and improving this paper.

Richard Dee, Sun Fellow
Jim Cates, Director, Tape Drive Development, Sun StorageTek
Ian Del Blaso, SAM-FS Marketing
Margaret Hamburger, SAM-FS Marketing
Jason Kranitz, Account Executive
Chuck Wenner, Systems Engineering Manager
Scott Matoon, Sr. Systems Engineer, Western Region
James E. Brennan, Senior Systems Engineer

Copyrights, Trademarks, etc.

Appendix A: Assumptions and Methodology

The assumptions and methodology for the included Tables are listed below.

The purpose in disseminating this document is to assist Industry in the investigative process by gaining insight and input from those in the field who have participated in the development of the model, or who may be interested in expanding it.

It is important to remember that this is an investigative work-in-process. Pricing and costs are shown only to provide a relative cost comparison with film-based archiving, and to show that a digital archive is, indeed, financially feasible today. Prices are not meant to be an offered quote or offer to sell of any kind.

The methodology reflected in this Digital Content Archive model is as follows:

1. Two libraries exist in geographically separated locations for disaster recovery (DR) purposes. Both libraries are identical in drives, slot count and front-end compute and disk configurations.
2. A 100TB Archive Object is ingested to the archive at the primary location.
3. At ingest read-after-write is implemented in order to ensure accuracy of data write, and the ECC is created and stored.
4. On read-back after write, the ECC is compared to ensure that the Archive Object sent to tape is, indeed, the bit-wise replicate of the Archive Object recorded on the tape.
5. A hierarchical storage manager creates four tape copies of the Archive Object ingested on disk and places them as below:
    a. Two copies remain in the primary library
    b. Two copies are sent to the secondary (DR) library.
6. All tapes are re-read and audited for data validity approximately every six months to monitor degradation. Tapes found to have bit errors are rejected and a new copy made from a verified valid copy in the library ensuring two valid copies always exist in each library. These steps are done automatically via the HSM, reducing human involvement in overall media maintenance. If migration is required (file format, tape software, media, etc.), migration occurs at this time.
7. Should both copies in one library fail an audit, two new copies can be made from a determined good copy in the DR library and then these tapes transferred to the first library. The intent is always to have two good copies in each library.
8. Content is copied to new tape media with the above validity checking every five years, discarding tapes older than that.
9. As additional library slots are needed they are added in increments of 1500.
10. Vacant slots are kept in the library to ensure space for blank cartridges to be used when new copies are required.
11. Tape drives are assumed to reach end-of-life and are replaced in five-year intervals.
12. When an Archive Object is re-written it is written to new and unused then-current-density cartridges; media are *not* reused in the Archive.
13. Disk and compute front-end infrastructures are replaced every five years.
14. Each library is assumed to reach end-of-life at twenty-year intervals and is replaced.
15. The model assumes a constant price of drives, tape media and libraries over time under the assumption that as technology drives down the price, added features and inflation will keep it constant.