Long term data integrity for large Audiovisual archives

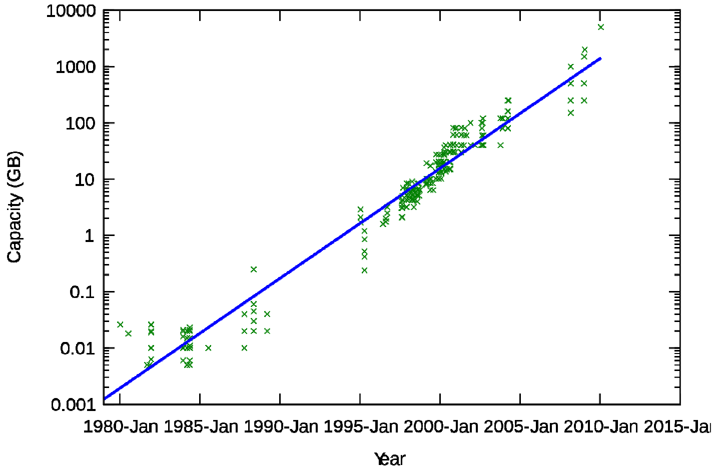
JTS 2010

Oslo

Matthew Addis, IT Innovation Centre
mja@it-innovation.soton.ac.uk

This presentation describes some of the work done in the PrestoPRIME project on how to achieve high levels of content safety when using IT systems for digital archiving and preservation. PrestoPRIME is a European Commission supported collaboration between broadcasters, archives, libraries, technology providers and researchers with the aim to develop new technology for digital audiovisual preservation and access.

Trend: increasing HDD capacity



- Doubles every 18 months
- 100 times every decade
- 1 million times every 30 years

© IT Innov

Many of you will have seen graphs like this that show the fantastic rate at which digital storage media increases in capacity, for example for hard drives the capacity doubles every 18 months, and has done so for the last 30 years – a million fold increase.

If that seems surprising, then think back to the 1980s and the first PCs where memory was in kilobytes and storage in megabtyes. Today the equivalent PC has terabyte storage and gigabyte memory.

In another 30 years, at this rate, and there is no reason to expect it won't be achieved one way or another, then you'll get an Exabyte of data on a single storage device – that's 1 million hours of uncompressed 2k film.

Trend: increasing recording density

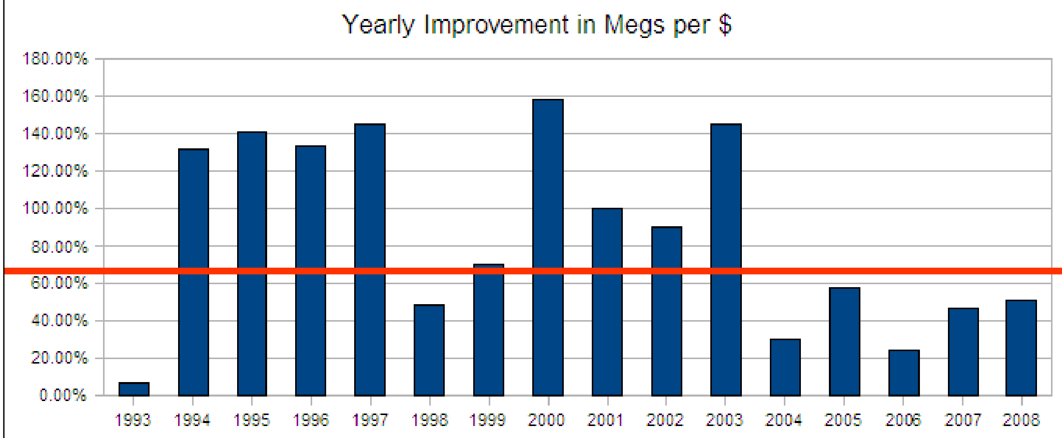
type of medium	audio data medium	recording capacity (minutes per square meter)
analog	6.35 millimeter wide 190.5 millimeters per second reel-to-reel magnetic tape	13.8
analog	33-1/3 RPM vinyl album	411
analog	90-minute audio cassette	184
digital	compact disk (CD)	8,060
digital	60-meter digital audio tape (DAT)	500
digital	2 terabyte 89-millimeter hard drive	4,680,000

http://www.americanscientist.org/issues/pub/2010/3/avoiding-a-digital-dark-age

© IT Innovation Centre and other partners of the FP7-ICT-2007-3 231161 PrestoPRIME consortium

And this increase in capacity is of course one of the things that makes IT storage ever more attractive for AV archiving.   You can see in this example, that you can already get thousands of hours of audio content on a single hard drive today.

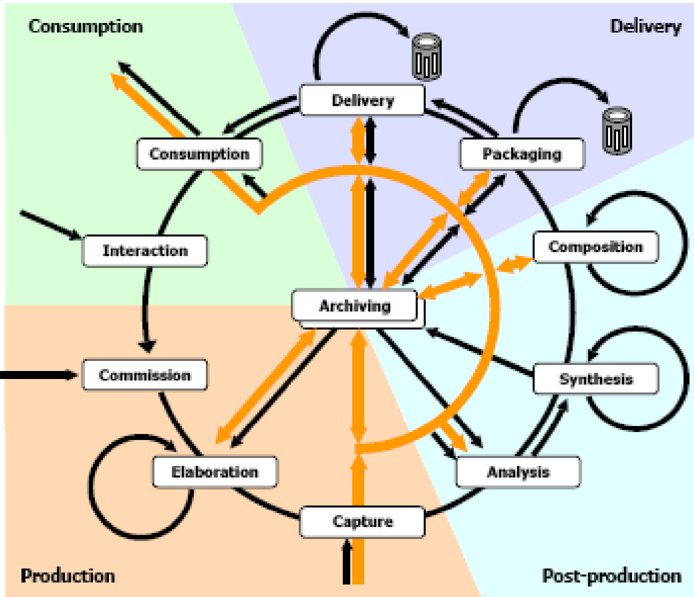
Trend: storage cost improvement



http://www.mattscomputertrends.com/harddrives.html

© IT Innovation Centre and other partners of the FP7-ICT-2007-3 231161 PrestoPRIME consortium

And this increased capacity doesn't come at increased cost. The unit price of storage is, bar a few wiggles, unchanged from year to year – which simply means ever more storage for your money.



- The ‘industry’ is going ‘tapeless’
- Networked sharing of file-based content

© IT Innovation Centre and other partners of the FP7-ICT-2007-3 231161 PrestoPRIME consortium

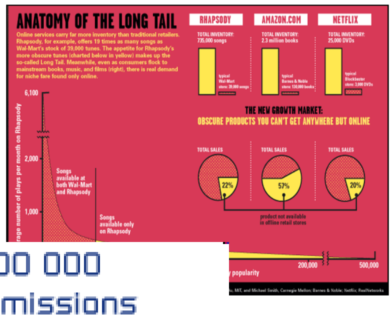
Or you may have seen pictures that like this, which shows the changing role of the archive in the production, post-production and distribution process. As the industry goes ‘tapeless’, which means working with files transferred over networks, the archive becomes much more central and embedded in the process.

This means production technology, archive technology and IT storage and network technology are all starting to blend together. And this makes a big difference to the accessibility and reusability of archive content in a professional context.

Access, access, access

creative commons

YouTube Broadcast Yourself



Waisda?

ina 100 000 émissions radio télé

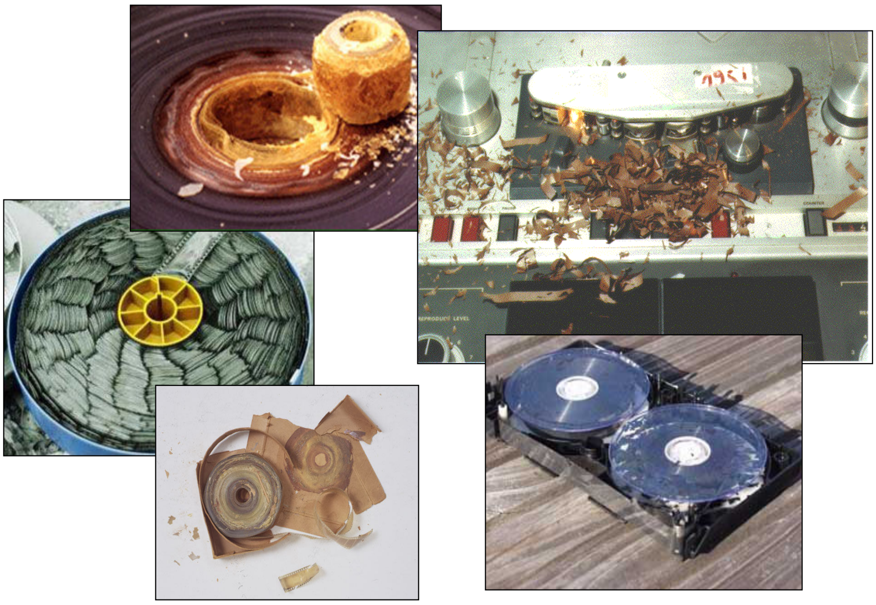
This is 4oD
The best of Channel 4, on demand

BBC iPlayer

SEVENTH FRAMEWORK PROGRAMME

© IT Innovation Centre and other partners of the FP7-ICT-2007-3 231161 PrestoPRIME consortium

Which is also matched by a drive for public access, and the interesting new models that this in turn enables for archive sustainability and enrichment (which others will talk about later).

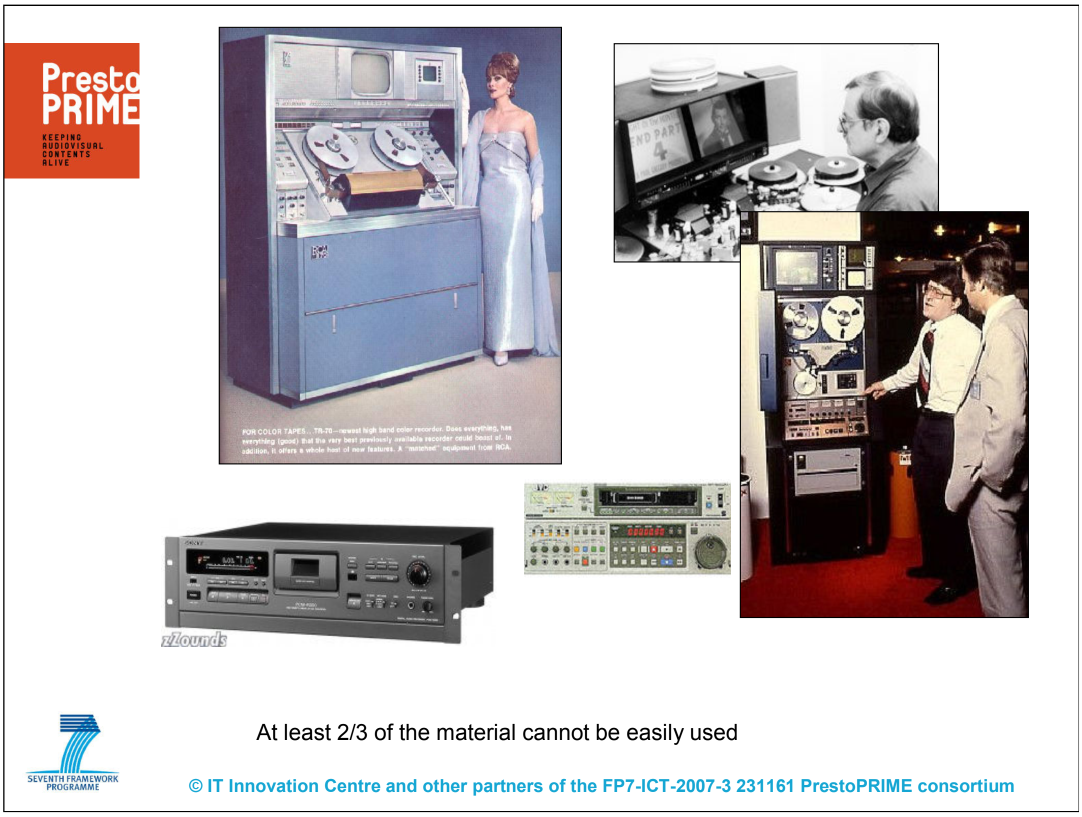


1/3 of material has deterioration

1/4 of material cannot be released as it is too easily damaged

© IT Innovation Centre and other partners of the FP7-ICT-2007-3 231161 PrestoPRIME consortium

Or you might be faced with the challenges of existing content, with a multitude of ways in which carriers can degrade or become fragile to use – which can apply just as much to more recent digital formats, e.g. digital video tape as it does to analogue carriers.



And if deterioration isn't a problem then sooner rather than later technical obsolescence will be. And it's this combination of deterioration, fragility and obsolescence that puts major fractions of our AV record at risk and forms the driver for mass digitisation or transfer projects in archives around the world, particularly for video, if not so much for film.

Benefits of IT storage and systems

- Lower cost
- Less space
- Easier access
- Fewer staff
- No more big manual migration projects
- Preservation moves closer to production

© IT Innovation Centre and other partners of the FP7-ICT-2007-3 231161 PrestoPRIME consortium

All these things: the promise of lower costs, less adminstration, easier access etc. along with new opportunities for example to capture and preserve content much earlier in its lifecycle, including essential technical and descriptive metadata, are all attractive reasons for using files and IT systems for audiovisual archiving

- But how safe is it?

© IT Innovation Centre and other partners of the FP7-ICT-2007-3 231161 PrestoPRIME consortium

But how safe are they? What guarantee is there that what you put in today you'll be able to get back out in 50 years time.

And if you can get it back out, how closely will it match the original – in several senses, including the 'bits' but also its visual or audible representation?

Trend: obsolescence

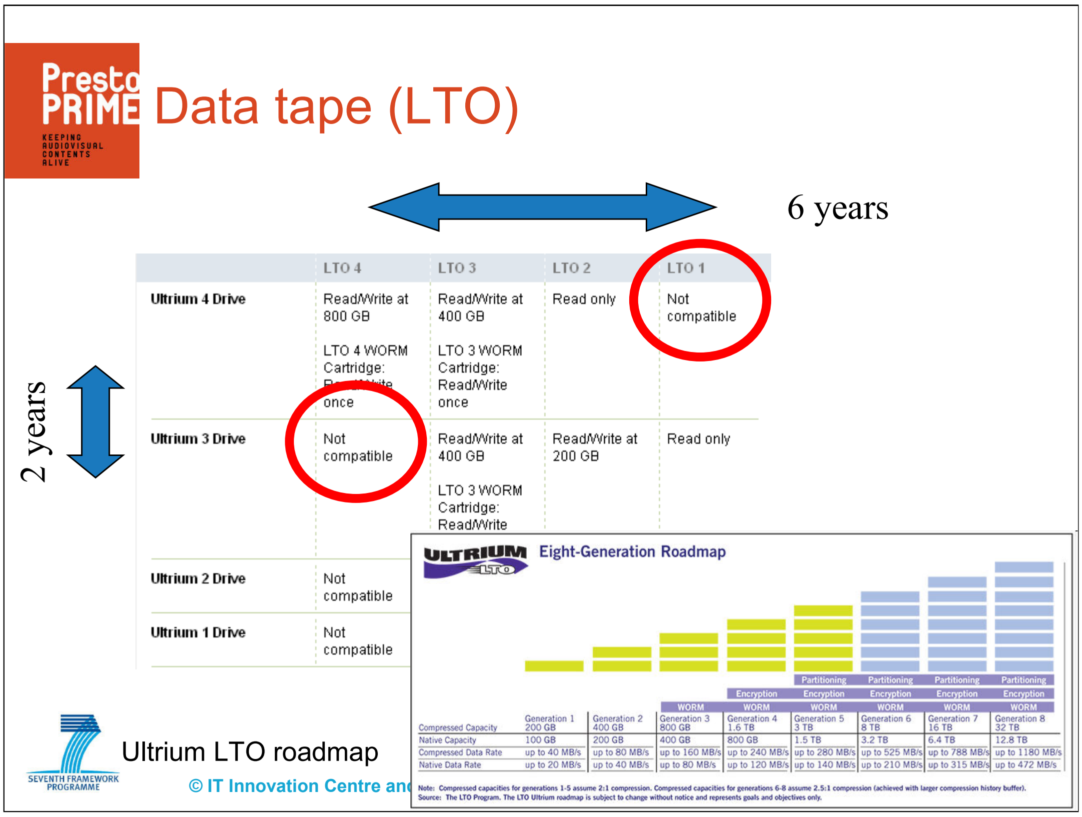
Medium	Storage Density bits/cm²	Life, years
Stone	10	10000
Paper	10^4	1000
Film	10^7	100
Disc	10^{10}	10

- Each change in ‘technology’ is 1000 times denser
- But the media lasts 0.1 times as long

© IT Innovation Centre and other partners of the FP7-ICT-2007-3 231161 PrestoPRIME consortium

So I showed earlier the trend for increased capacity. But this trend is also accompanied by shorter lifetime, not just for entire technologies, but also generations of that technology.

Media deterioration is not so much an issue as perhaps in the past, because player obsolescence will probably get you first.

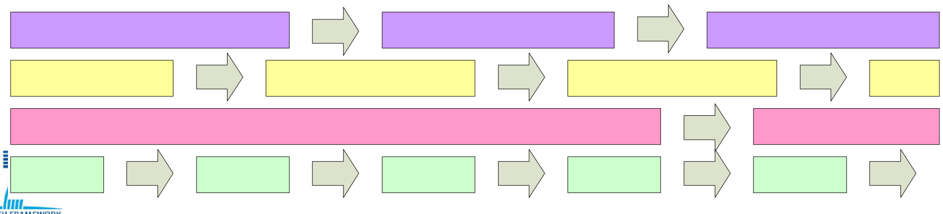


And a good example of this is LTO data tape. With the recent announcement of LTO5 and an extension to the LTO roadmap, data tape as a technology continues, but when you look at the details of backwards compatibility, then you see that obsolescence is swift.

With each new generation, every two years or so, comes twice as much storage capacity, but not the ability to use the new media in older drives. After a couple of generations its not possible to write old media in new drives. Another generation after that and old media can't even be played any more.

Obsolescence is everywhere in IT

- Encoding formats
- Media formats
- Storage hardware
- Operating systems
- Management software
- Networking



SEVENTH FRAMEWORK PROGRAMME

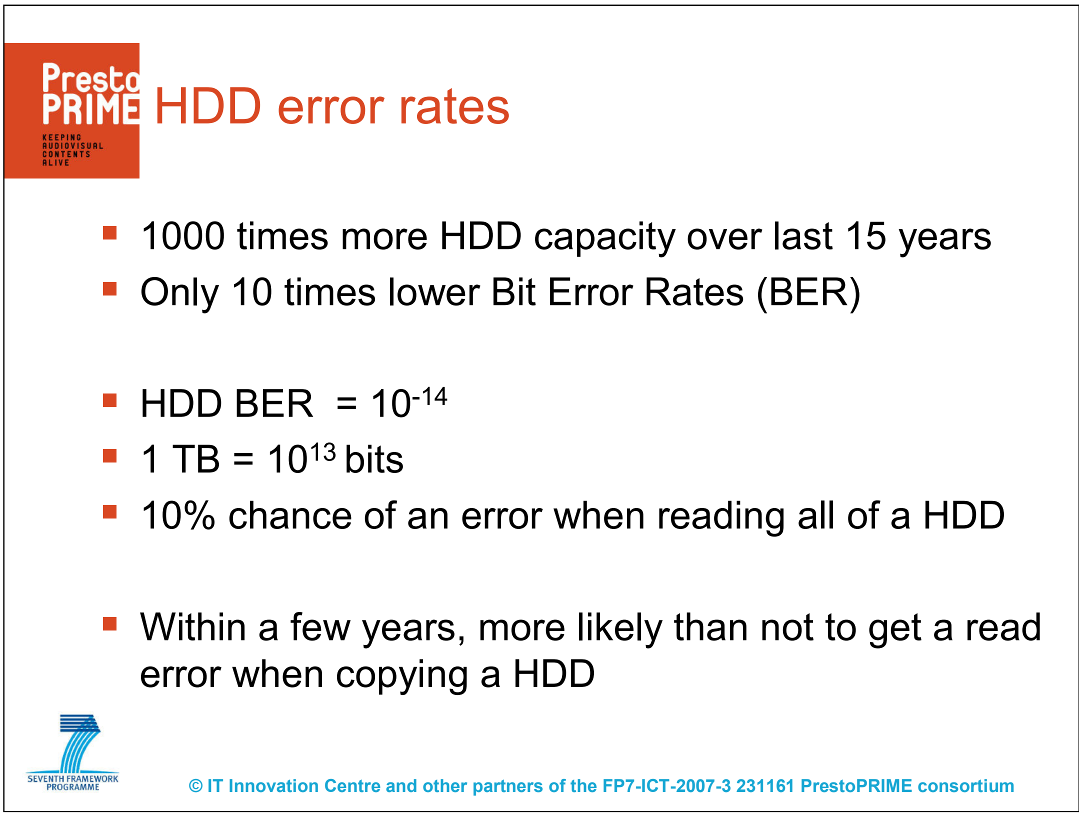
© IT Innovation Centre and other partners of the FP7-ICT-2007-3 231161 PrestoPRIME consortium

And this sort of obsolescence is everywhere in IT – at all levels of the technology stack – and with differing timescales.

The result is the need for ongoing technology migration, which for any scale of AV content requires automated systems.

The risk here is that doing nothing or taking your eye off the ball for even a few years puts content at risk.

This is very much in contrast with the 'items on shelves' approach for analogue media where 'doing nothing' for 10 or 20 years, other than using controlled storage conditions, would still give a good chance of being able to play content back.



If we now look at error rates of the media itself, then this isn't anywhere near keeping pace with the rate at which capacity is increasing. A modern hard drive might have a bit error rate of 1 in 10 to the 14. This means you could expect to get some form of error, even if it's just one bit, every time you read 10TB of data from a hard drive. This is tiny. It has to be said that hard drives are fantastic pieces of engineering.

But, with increases in bit rates, resolutions, sampling etc. for AV formats, the number of bits in a file is now huge.

We're getting to the point that it is more likely than not to encounter an error when reading all the data from a hard drive. This has big consequences for both storing large AV files, and also on the use of hard drives inside systems that take further steps to prevent these errors.

HDD lifetime

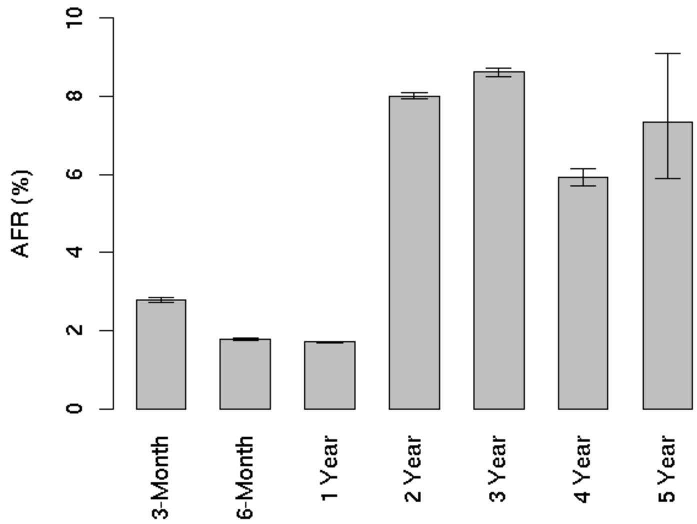
- Manufacturers say:
  'Mean Time Between Failures' = 1 million hours
- What does a MTBF of 1,000,000 hrs mean?
- What is does ***not*** mean:
  - A HDD will typically last 100 years
  - Or, the failure rate is 1% each year
- Lifetime of a HDD is 3-5 years

© IT Innovation Centre and other partners of the FP7-ICT-2007-3 231161 PrestoPRIME consortium

And then there's the chance of the hard drive as a whole failing. Here manufacturers talk about Mean Time Between Failures of a million hours, which is about 100 years. But whilst this sounds good, this is a fairly useless statistic and gives a false impression unless interpreted very carefully. It certainly doesn't means that you can expect a hard drive to last 100 years!

Its commonly accepted that the useful lifetime of a hard drive is up to 5 years. Use it beyond this and you risk increased failure rates as well as technical obsolescence.

HDD failure rates



Google study of Annual Failure Rates in HDD servers

© IT Innovation Centre and other partners of the FP7-ICT-2007-3 231161 PrestoPRIME consortium

Field studies, e.g. by Google or NetApp, involving hundreds of thousands of drives in real world systems reveal the real failure rates of hard drives in their earlier years. And this is what matters especially if you have any designs on a 'hard drives on shelves' archiving policy.

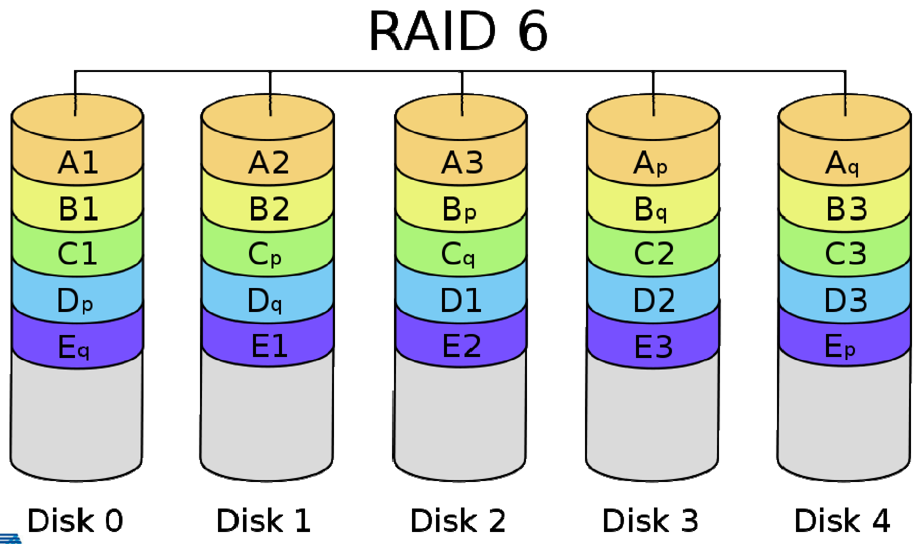
Failure rates of hard drives can easily be 5% per year. Interestingly,that isn't anywhere near as dependent as you might think on how often the drive is used or the temperature at which it operates.

Indeed, there is some evidence that cooled and infrequently used drives have higher problem rates than those that are continually spinning in room temperature servers.

Not using a drive, i.e. keeping it unpowered for long periods of time, e.g. years, has historically caused many problems, e.g. sticking heads, and even with modern drives is not something that is 'designed in' by the manufacturers – so whilst there is limited information on failure rates in these circumstances, it would be reasonable to assume for now that failure rates for 'drives on shelves' is likely to be high.

And the thing to bear in mind is that if the drive fails, then all data is potentially lost unless very expensive recovery operations are undertaken.

The IT Industry knows this already



© IT Innovation Centre and other partners of the FP7-ICT-2007-3 231161 PrestoPRIME consortium

But of course the IT industry knows this already. This is why approaches like RAID exist that combined multiple drives into one storage system to counter drive failures in whole or in part.

But systems bring their own problems

“Disk failures are not always a dominant factor of storage subsystem failures, and a reliability study for storage subsystems cannot only focus on disk failures. Resilient mechanisms should target all failure types”

2008 NetApp study of 1.8M HDD in 155,000 systems

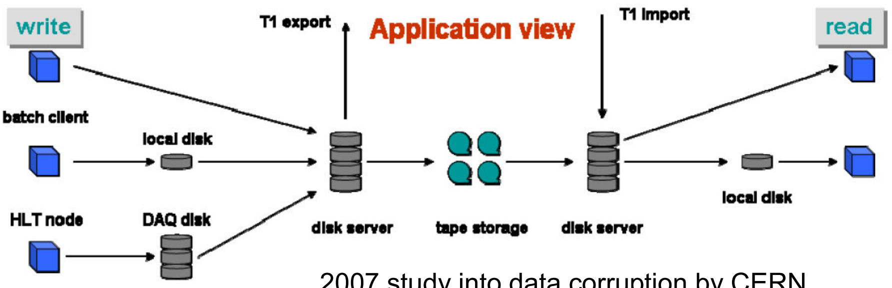
© IT Innovation Centre and other partners of the FP7-ICT-2007-3 231161 PrestoPRIME consortium

But these extra systems bring with them extra complexity and new bugs/errors and ways to lose or corrupt the data within them – despite the best intentions of their designers and builders.

The horror of ‘bit rot’

David Rosenthal’s blog
http://blog.dshr.org/

- Errors can be silent (latent)
  - Permanent and undetected corruption of data
  - Deeply worrying for archives
  - Seen in field studies (if you know how to look)



2007 study into data corruption by CERN

© IT Innovation Centre and other partners of the FP7-ICT-2007-3 231161 PrestoPRIME consortium

Field studies of IT storage systems, including those engineered specifically to avoid loss, e.g. using RAID, error correcting memory, resilient data transfer protocols etc. show data corruption is a fact of life. An example is a study by CERN which showed data corruption rates as high as 1 in 10 to the 9 – actually worse than for individual drives – which reflects the wide array of problems that can occur in a complete system as opposed to just one part of it.

Most worrying is that this loss was silent and permanent. You only see it if you know it can happen and then choose to look for it. Sometimes this is called ‘bit rot’. For lots more on this topic I’d suggest looking at David Rosenthal’s blog for a regular round up of this area. Basically, reliability of IT systems is orders of magnitude short of the level needed to be considered ‘safe’ from a preservation perspective.

Cost of reducing chance of loss

- Storage capacity increasing very quickly
- Storage speed and error rates not keeping pace

- Increasingly complex measures needed
- Disproportionate time and cost needed to manage integrity

RAID	Scrub	Sector Checksum	Block Checksum	Parent Checksum	Write-Verify	Physical ID	Logical ID	Version Mirror	Chance of Data Loss
√									0.602%
√	√								0.602%
√	√	√							0.322%
√	√		√						0.041%
√	√			√					*0.486%
√				√					*0.153%
√	√		√		√				0.002%
√	√		√			√			0.038%
√	√		√				√		*0.033%
√			√				√		*0.010%
√	√		√			√	√		*0.031%
√			√			√	√		*0.010%
√	√		√				√	√	*0.004%
√			√				√	√	*0.002%
√	√		√			√	√	√	0.000%

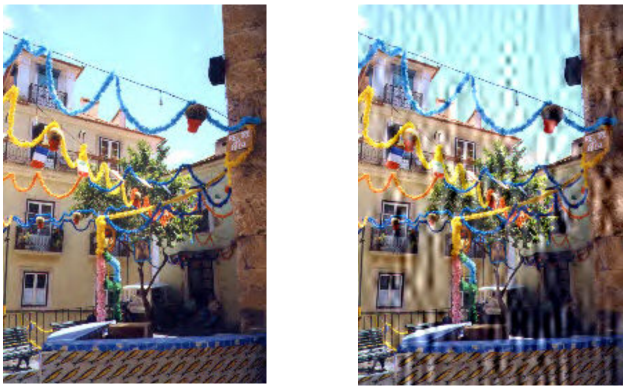
Table 3: **Probability of Loss or Corruption.** *The table provides an approximate probability of at least 1 data loss event and of corrupt data being returned to the user at least once, when each of the protection schemes is used for storing data. It is assumed that the storage system uses 4 data disks, and 1 parity disk. A (*) indicates that the data loss is detectable given the particular scheme (and hence can be turned into unavailability, depending on system implementation).*

© IT Innovation Centre and other partners of the FI

And the more efforts you make to lower chance of data loss, then the more complex and expensive the system becomes.

Striving for perfection results in massive and unsustainable cost. Better is to accept that loss will happen and balance cost with lowered risk, then find the most acceptable compromise, i.e. a 'cost of risk of loss' approach.

Cost of not reducing loss (1)



Volker Heydegger study on file format sensitivity to corrupton

JPEG2000 with one error per 100KB

- Compression = Corruption amplifier
  - Corrupting 0.001% of encoded image results in 30% of pixels affected in decoded image

© IT Innovation Centre and other partners of the FP7-ICT-2007-3 231161 PrestoPRIME consortium

And this is the cost of not reducing the loss.

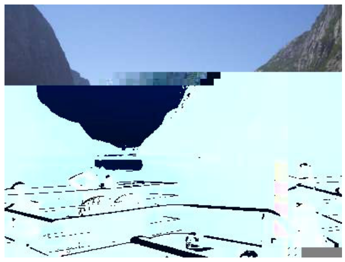
These are two JPEG2000 images, which is of course very relevant with JPEG2000 emerging as a candidate preservation format, especially in its lossless form.

There is a one byte error in the file for each image. Depending on where that error is, it can have major consequences on how useable the image then becomes. Volker Heydegger from University of Köln has done some great work looking at the robustness of images to data corruption.

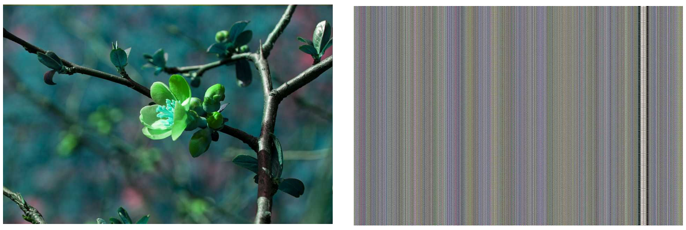
The interesting thing is the way the use of compression can amplify the effects of data corruption, which applies just as much to lossless compression - this isn't an escape option.

Cost of not reducing loss (2)

BMP file with one error per 256 bytes (1400 errors)



GIF file with a single error (in 14 KB)



1 byte error in TIFF image with (a) no compression, (b) zip compression

© IT Innovation Centre and other partners of the FP7-ICT-2007-3 231161 PrestoPRIME consortium

And neither is other encodings an escape route either, be they image specific e.g. GIF or generic, e.g. zip. Indeed, CERN found corrupting 1 byte in a range of zip files resulted in 99% of them becoming unusable.

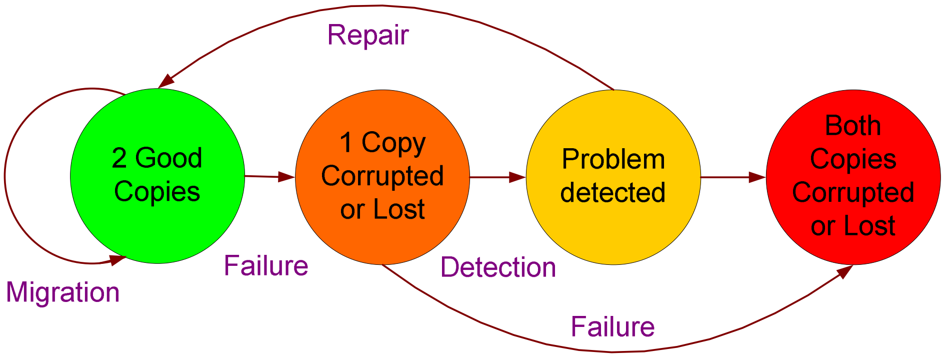
Since most preservation formats for visual content are essentially a set of images, i.e. intra-frame encoded and not inter-frame encoded, then any video codec used in preservation is likely to have this problem – although further research is certainly needed.

The results are equally horrific for compressed audio content too.

Only uncompressed formats show any ‘graceful’ behaviour when data within them is corrupted.

Managed 'cost of risk of loss'

- Multiple *independent* copies
- Detection and correction of failures
- Migration to address obsolescence
- All activities have a cost



So, having looked at some of the issues of IT technology reliability and longevity and the consequences on content, the question is what to do about it.

Here there are plenty of options, so the challenge is one of how to identify and apply the most appropriate ones.

Looking at the diagram, a commonsense approach is to keep multiple copies of content, typically using different technologies and in different locations, and then migrate the technology stack for each so the copies remains useable. Now, in keeping these copies there is always the chance that for one reason or another one of the copies is damaged or lost. This is shown in orange.   This represents some form of failure in the system.   But it's only after this problem is detected, shown in yellow, that any action can be taken, e.g. to repair or replace the damaged or lost copy. And, in this two copy example, if at any time something happens to the second copy after the point that the first copy has a problem that isn't rectified, then there is a risk that content is permanently lost or damaged – shown here in red.

Clearly the faster that failures can be detected and repaired then the lower the overall risk of loss.

But this has a cost. All activities have a cost, including migration. So, the approach is to look at all these costs to find the best solution.

Approaches

- Use **longer lived** storage technology
  - E.g. Printing bits to film
- Use **more reliable** storage technology
  - E.g. data tape instead of HDD on shelves
- Make **more copies**
  - E.g. off site deep archiving
- Encode so **content is more resilient**
  - E.g. Graceful degradation
- Use **concealment**
  - E.g. Interpolation to replace corrupted frames or blocks
- **Check often** and **fix quickly**
  - E.g. scrubbing of HDD servers

And here's some of them. I'll cover several in detail. But they all come down to how often you need to migrate or check content, and how often you need to repair.

However, like a squishy balloon, if you nail down one area, then another can get worse.

For example, if you use compression then you can make more copies for the same storage cost, but each copy is more sensitive to data corruption and is harder to repair.

Long lived digital media

- Preservation grade DVDs
- Magneto Optical disks
- Digital Film
- Rosetta Discs

- Can be very expensive
- Lock in to a vendor or particular approach
- Not mainstream

- *Risk is the longevity of the vendor, not the technology*

SEVENTH FRAMEWORK PROGRAMME

© IT Innovation Centre and other partners of the FP7-ICT-2007-3 231161 PrestoPRIME consortium

Making the technology longer lasting, and hence reducing the need for migration and risk of obsolescence is an obvious option. Options include archival blu-ray, printing bits to film, or even more esoteric approaches e.g. the long-now foundation's rosetta disk.

The problem shifts however. It will typically becomes one of increased cost, especially relative to ever falling commodity IT alternatives, and in particular how long the vendor will last for – long-lived archive media rarely achieves mainstream industry.

It's all very well having content on a 50 year magneto optical disc, but what if the company that produces the disc goes bust then the content is effectively lost.

Data tape

- Relatively ‘safe’ technology (compared to HDD)
- Typical ‘problem rates’ are 0.1 – 1% of tapes

- Most problems from data tape come from drives
  - Malfunctioning or worn drives that damage tapes
  - New drives that don’t handle older generations properly

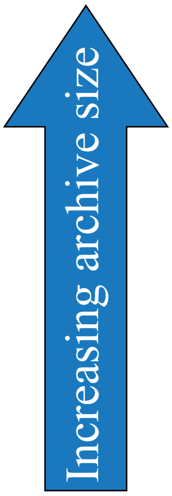
- *Field studies show no data lost where multiple copies have been made and integrity checked*

© IT Innovation Centre and other partners of the FP7-ICT-2007-3 231161 PrestoPRIME consortium

Or you can pick a more reliable technology, which reduces the need to check it so often and hence reduce the cost of ‘active measures’ or complicated systems.

Data tape is a good example, with field studies showing reliability and error rates that are orders of magnitude better than raw hard drives. Indeed, several large archives have already gone through multiple migrations of 10s or 100s of terabytes of content and have verified that they haven’t lost a single bit as a result. Mostly because they made sure that they had multiple copies, they checked data integrity every time data is moved and they automate tape management using robots.

Tipping points



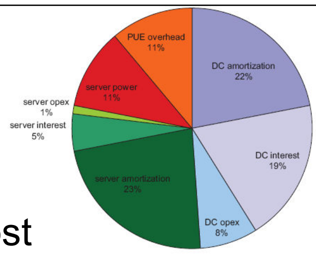
- All on tape (2 copies), hard disk only for staging
- Frequently used on hard disk, two copies tape
- All on hard disk (1 copy), safety copy on tape
- All on hard disk (2 or more copies)
- All on flash (2 copies, e.g. USB sticks)

© IT Innovation Centre and other partners of the FP7-ICT-2007-3 231161 PrestoPRIME consortium

But data tape tends to be more appropriate for the larger archives who can afford the initial costs of the drives and robotic systems. Much more tempting for the smaller guys is the use of hard drives – yet this is potentially a much less safe technology and is used by archives who are much less likely to be aware of the issues.

And what's 'large' and what's 'small' will change, with the cross-over and mixing of technologies occurring at different points. Here the risk is that an archive familiar with one technology and approach will find itself moving to another technology with different loss characteristics where the same safety techniques won't work anymore.

Make more copies



- TCO is a multiplier of unit media cost
- HDD storage is 2-3 times cost of data tape
- Real world costs halve every 2-3 years

- Take today's raw media cost and x10 for annual rate
- 1TB online for 1 year costs $1000
- Multiply again by 4 for lifetime cost

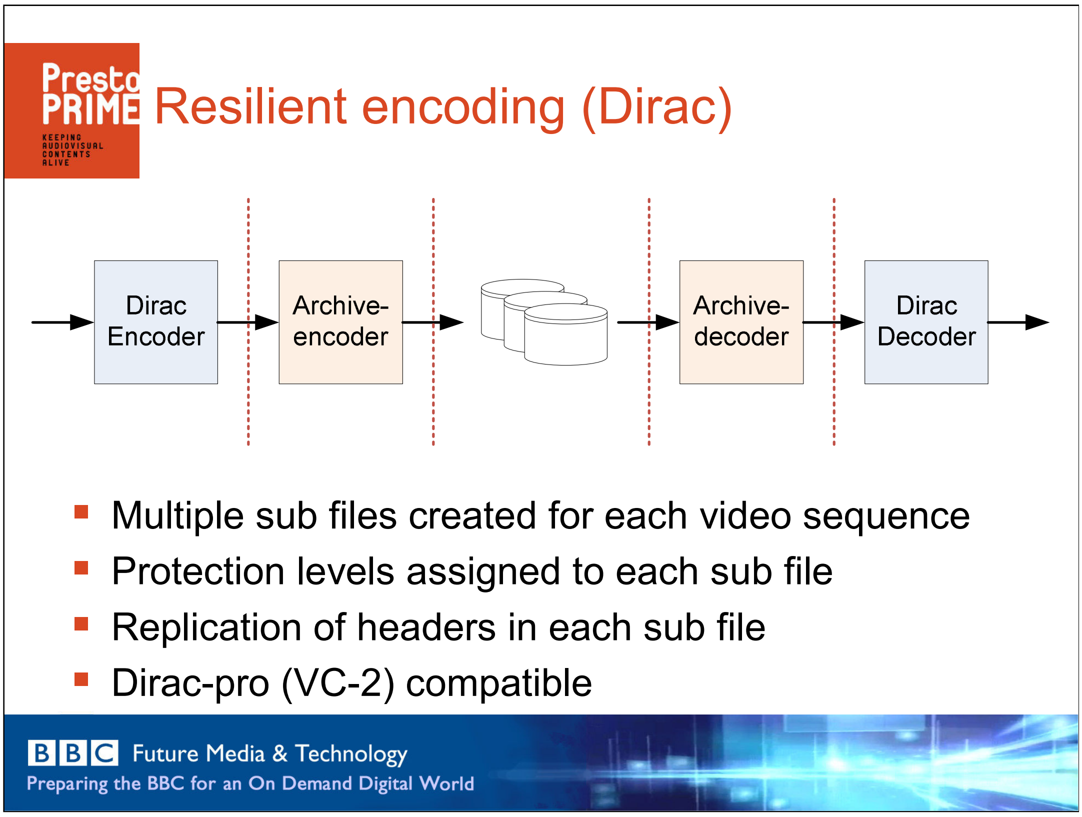
- £1 per hour of audio on data tape *forever*
- £10k per hour of 4k film on hard disk *forever*

© IT Innovation Centre and other partners of the FP7-ICT-2007-3 231161 PrestoPRIME consortium

Making more copies is another obvious answer, but only if you can afford it, with the total cost of ownership over time (people, space, power, cooling, maintenance, migration etc.) all adding up so that even with the falling cost of storage, the total cost can be massive.

Making more copies means you can get away with checking them less often, or, with a mix of technologies, you can worry less about obsolescence.

This is why compression is of course so attractive as it allows more copies for the same cost.

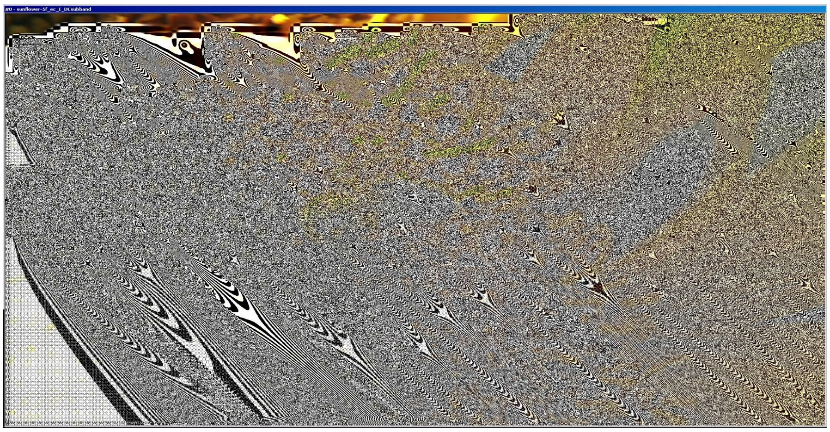


Or you can change the way the content itself is encoded.

The BBC are working on an archive version of their dirac encoding which is specifically designed to be more resilient to data corruption. Here a standard dirac encoded file is split apart into its constituent pieces, each of which is stored in a ‘sub file’ that is then replicated or stored in a way that matches its specific sensitivity to data corruption.

Example: sub file 1

- Errors on DC sub band data:



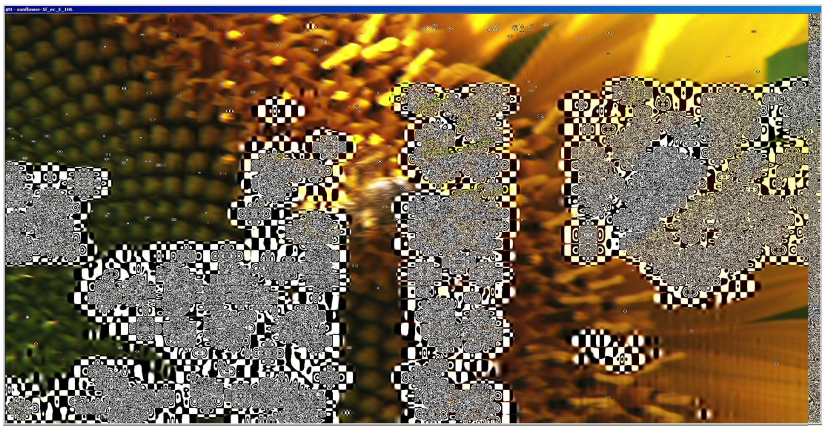
BBC Future Media & Technology
Preparing the BBC for an On Demand Digital World

So for example, maximum protection might go to the DC component since corruption here can cause catastrophic effects

Presto PRIME
KEEPING AUDIOVISUAL CONTENTS ALIVE

Example: sub file 2

- Errors on low freq sub band:

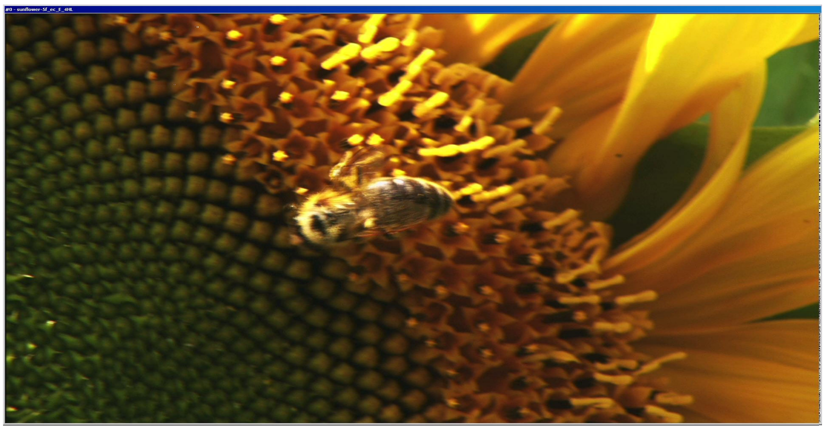


BBC Future Media & Technology
Preparing the BBC for an On Demand Digital World

And plenty of protection to the low frequency components too.

Example: sub file 3

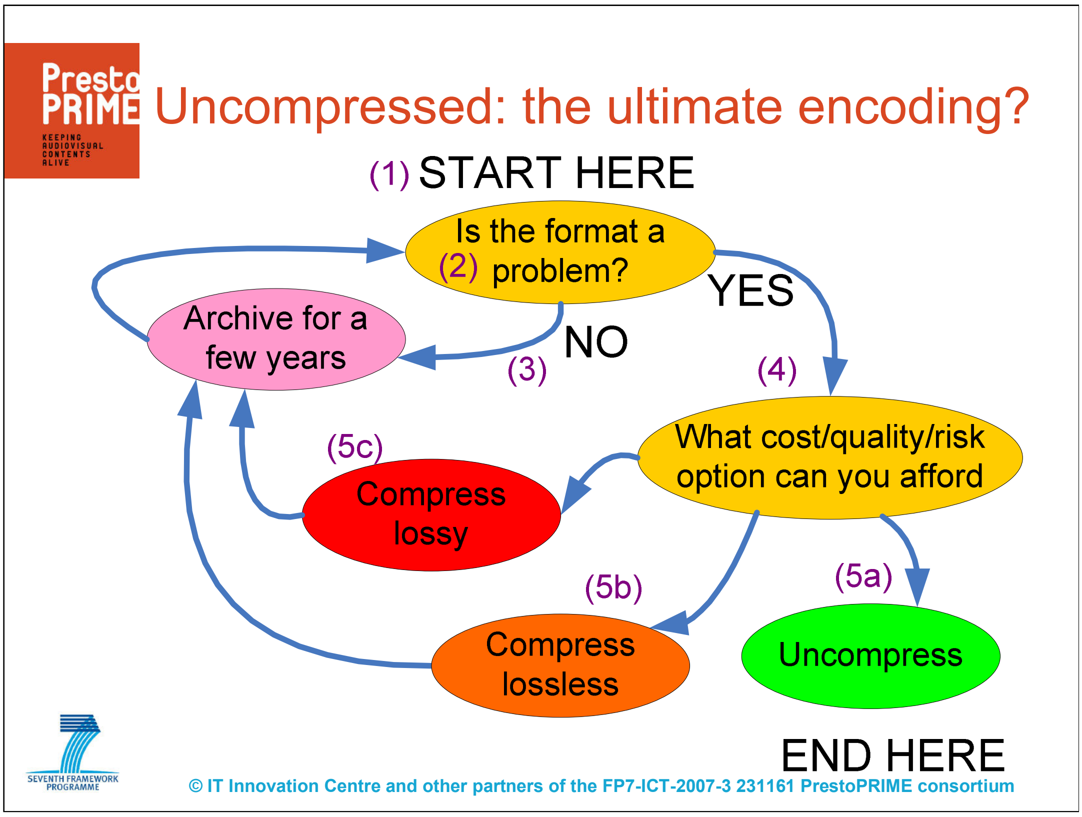
- Errors on high freq sub band:



BBC Future Media & Technology

Preparing the BBC for an On Demand Digital World

But corruption of the high frequency component has much less visual impact, so there is potential here for using less reliable and hence cheaper storage, or not to check and repair this part so often. In this way, copies of the content can remain usable for longer before repair needs to take place.



Of course, uncompressed could be considered as the ultimate encoding. It's simplicity and resilience through inherent redundancy makes it a relatively reliable and long-lived way to store content. It is less likely to become obsolete and less likely to be affected seriously by data corruption before it does become obsolete.

Some archives are already adopting this approach, e.g. the BBC in their D3 project where they are transferring from D3 tape into a file that contains the uncompressed SDI bitstream from the D3 player.

The problem is one of cost, which brings with it a preservation 'game' of using compressed formats in the short term to save cost and then moving to uncompressed when budgets allow. The rule here is use compression just once, since transcoding, especially between lossy formats, has possibility to introduce loss of quality.

Concealment

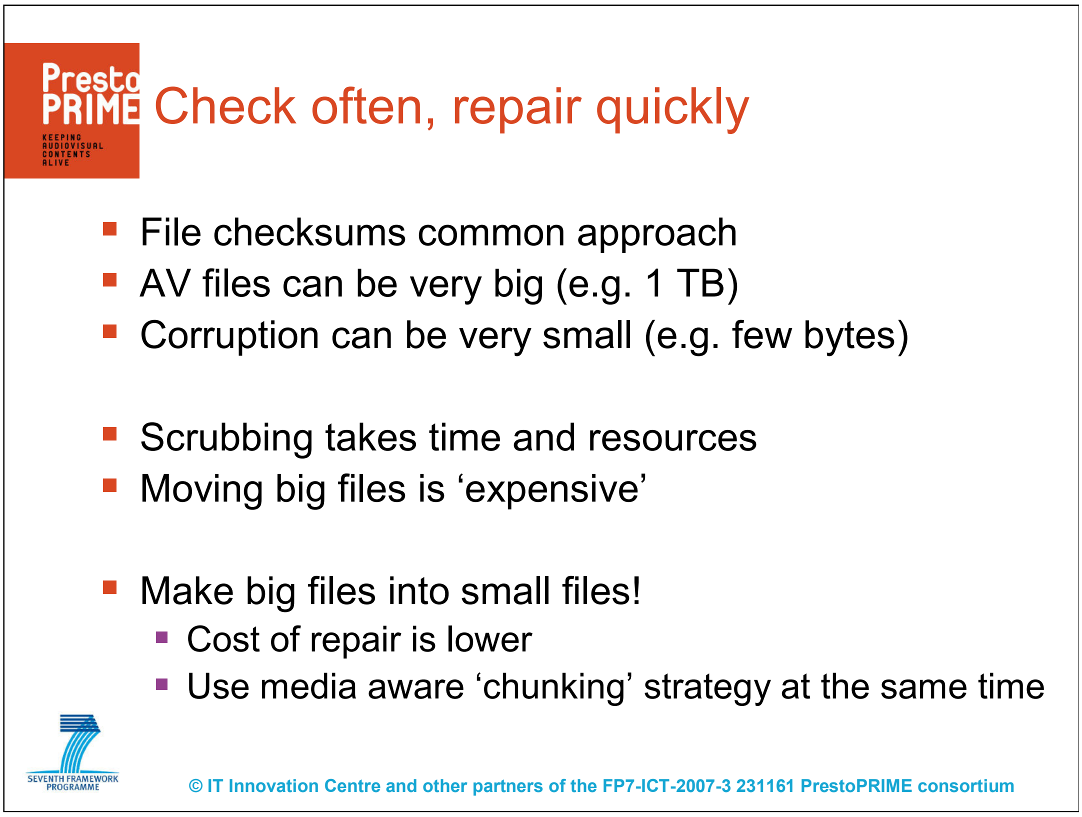
- Common approach in the days of analogue
- Used in digital video tapes, e.g. DV
- Tight coupling between encoding, carrier and playback
- No general purpose concealment technology for file based storage using IT systems
- Good area for further research?

© IT Innovation Centre and other partners of the FP7-ICT-2007-3 231161 PrestoPRIME consortium

And if errors do exist and can't be repaired completely, i.e. to give a bit-perfect version, then concealment is another option. Common for analogue, and also present in many digital video tape formats too, e.g. DV, this automatically conceals problems at point of playback.

This works because the carrier, encoding, error handling and concealment are all tightly coupled together in a single technology – something where there is no general purpose equivalent in the IT storage world.

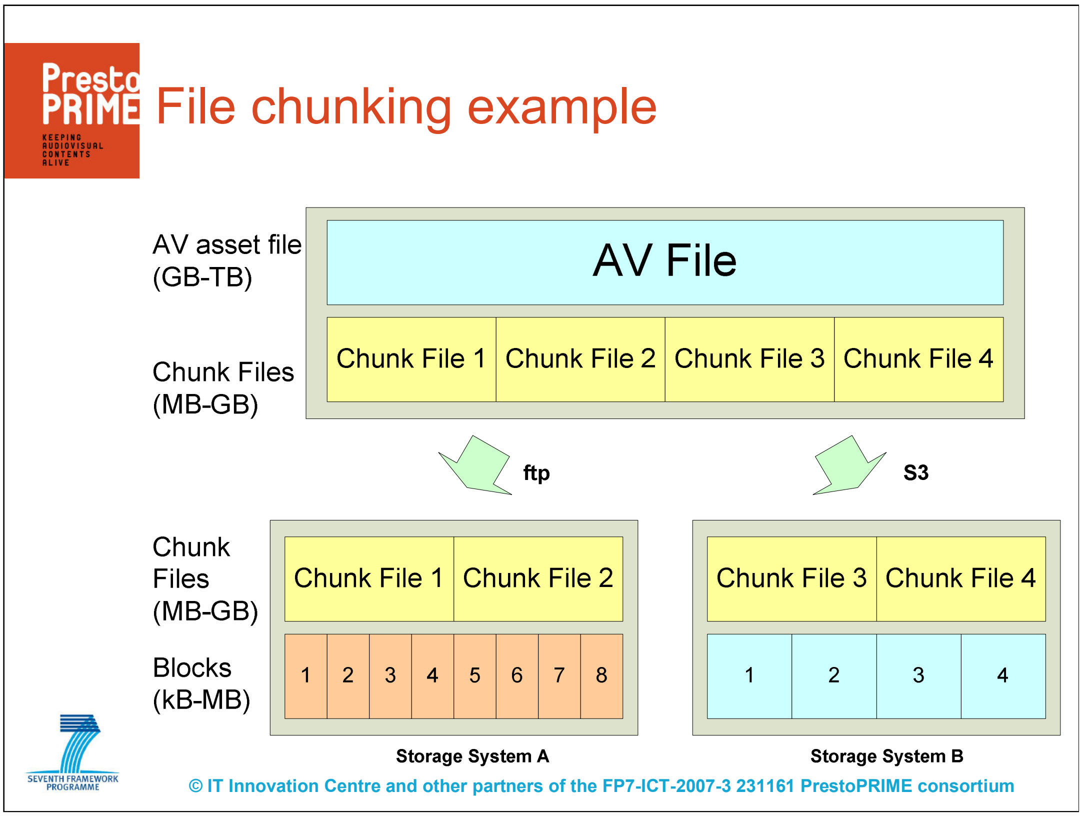
IT storage doesn't understand audiovisual files. Audiovisual coding isn't optimised for the errors in IT storage. Seems to me an area for more work.



Then finally there is the approach of checking files often and fix problems quickly, which is how RAID and scrubbing works etc. But you need to do this at a high level across all storage, networking, processing etc in the archive. This is why many archives already checksum their files.

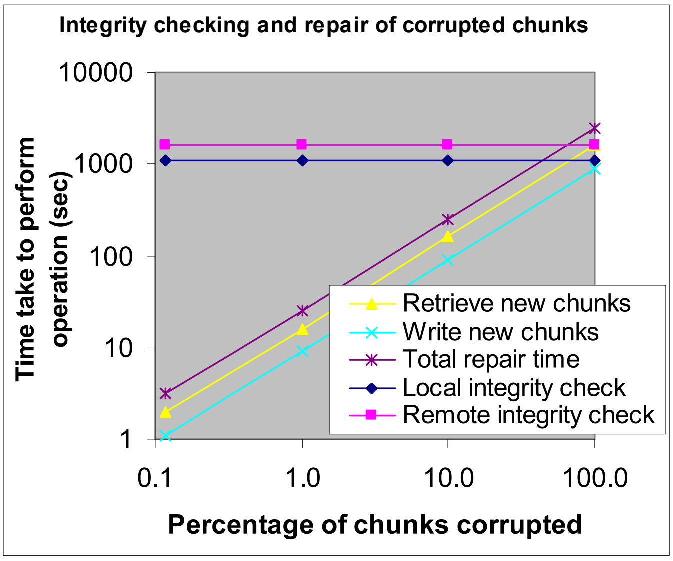
Problem is that the files are big and the errors are small, which means the cost of repair can be very high, e.g. retrieving massive files out of a deep tape archive just to fix a few bits in a local disk copy.

In a UK project called AVATAR we're looking at how to address this by making big files into smaller files so that end-to-end integrity management more efficient.



So big files get chopped up into smaller ones which are then replicated and distributed to different locations

Chunking reduces repair costs



Repair of a 59GB file in 10MB chunks with varying levels of corruption

© IT Innovation Centre and other partners of the FP7-ICT-2007-3 231161 PrestoPRIME consortium

If the checksum for a chunk of a big AV file fails in one location, then only the corresponding chuck needs to be copied to replace it from another location.

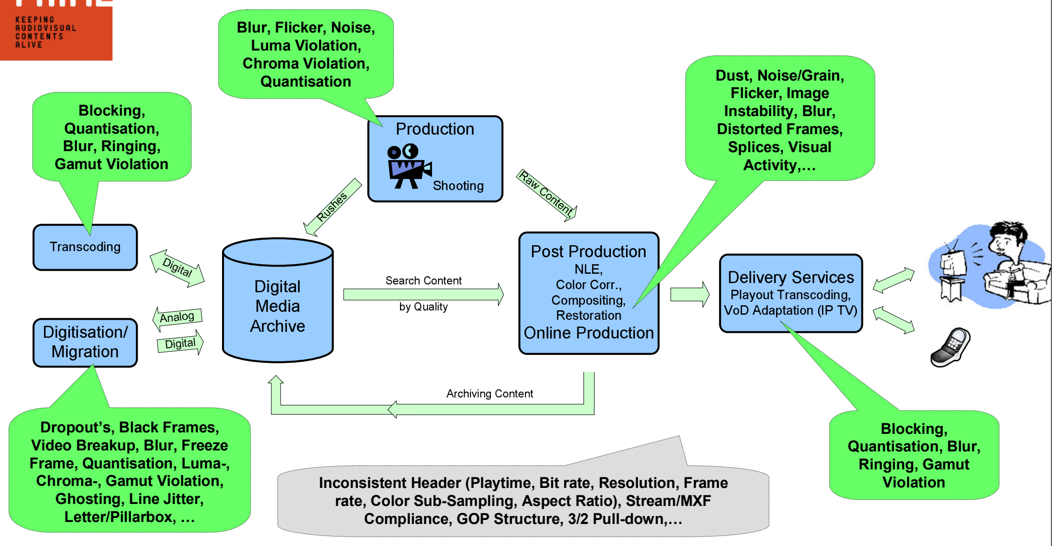
There's more to life than bits

- Purely technical checks can detect:
  - stream/format compliance, storage errors, transfer errors
- But these do not "see" errors in the video.
- Need Content-based Video Quality Analysis
  - Try to "interpret" the video material…
  - Detection of severe visual defects

As mentioned, some AV archives already do basic checks at the syntactic or bit level on content, e.g. compliance to MPEG standards or use of MD5 checksums on a bitstream. But it is perfectly possible to meet all these checks and yet still have content with problems at the visual level.

In PrestoPRIME, Joanneum Research are working on how to detect a range of quality problems, e.g. video breakup, blocking, black frames etc. so archives can apply more extensive quality assurance.

Essence: video quality assessment

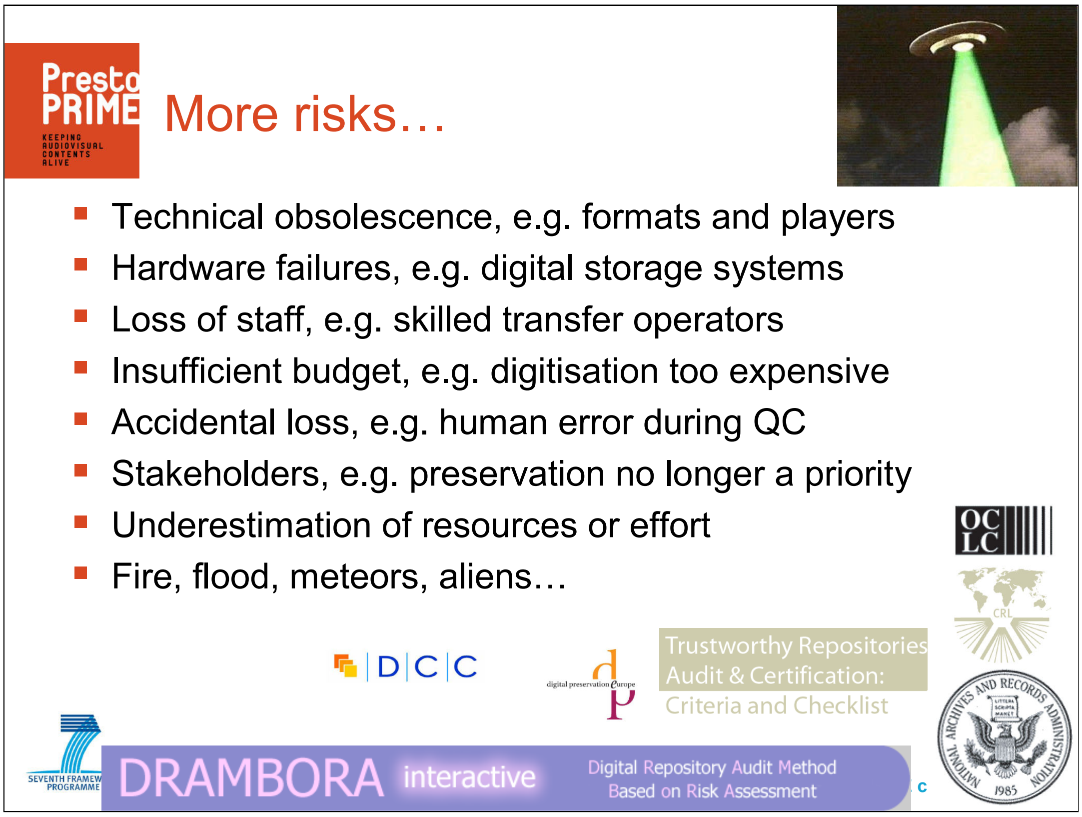


Need quality assessment throughout the preservation process: selection, ingest, migration, access

© IT Innovation Centre and other partners of the FP7-ICT-2007-3 231161 PrestoPRIME consortium

And this quality assessment is particularly important given the opportunity for video problems to creep into content at all stages in the content lifecycle, including its original production, its transcoding and delivery, and also during any subsequent manipulations, e.g as part of content reuse.

So, bit or file level checks are of course necessary, but so too is proper content quality checking.



And risks to content quality are just one example of a wider set of risks that come from file-based working and IT systems, or digital preservation in general.

So in PrestoPRIME we're investigating this and have recently used a risk assessment methodology based on DRAMBORA from the trusted repository world and OCTAVE from the information security world to look at the risks to content from IT systems.

Files: 37 risks from ‘IT’

- Risks of **loss of data authenticity and integrity**
  - Loss of ability to track and record what’s been done
  - Changes to integrity or authenticity go unnoticed.
- Risks of **data destruction or degradation**
  - Loss or corruption of data
  - People: deliberate or accidental damage
  - Technology: bit rot, obsolescence
- Risks to data through **loss of services**
  - E.g. loss of routine integrity checks
  - Loss or pressure on resources used to do preservation
- Risks to through **mismatch of expectations**
  - Service providers don’t meet archive needs

© IT Innovation Centre and other partners of the FP7-ICT-2007-3 231161 PrestoPRIME consortium

These risks include all the issues of data corruption in imperfect IT storage systems that I’ve already presented, but also risks to content from not maintaining enough capacity to perform preservation actions, which is always a challenge when there is a strong need to access content and this uses shared resources.

Further risks come from using service providers, which obviously means third-parties, but could equally be IT service provision within an organisation, where lack of clear agreements or expectations can result in problems.

Example risks

Risk ID	Title	Example
R30	Hardware Failure	A storage system corrupts files (bit rot) or loses data due to component failures (e.g. hard drives).
R31	Software Failure	A software upgrade to the system looses or corrupts the index used to locate files.
R32	Systems fail to meet archive needs	The system can't cope with the data volumes and the backups fail.
R33	Obsolescence of hardware or software	A manufacturer stops support for a tape drive and there is insufficient head life left in existing drives owned by the archive to allow migration
R34	Media degradation or obsolescence	The BluRay optical discs used to store XDCAM files develop data loss.
R35-R38	Security	Insufficient security measures allow unauthorised access that results undetected modification of files.

SEVENTH FRAMEWORK PROGRAMME

© IT Innovation Centre and other partners of the FP7-ICT-2007-3 231161 PrestoPRIME consortium

So just to illustrate, this is what risks look like with some examples of their manifestation.

I won't go through these in any detail, because there's a big report that's available from PrestoPRIME that you can use to get all the details.

Loss of data authenticity and integrity (origins)

- Lack of, or failure to follow, proper process
- Failure to record all actions performed within the archive
- Failure of archive storage systems or processing of content
- Failure to record attempts (deliberate or otherwise) to breach systems
- Failures at remote storage service providers
- Deliberate attack by disgruntled employees
- Deliberate attack by hackers or other third-parties
- Failure of preservation systems to correctly apply preservation actions

© IT Innovation Centre and other partners of the FP7-ICT-2007-3 231161 PrestoPRIME consortium

And for each of the risks identified in the report, we look at where the threats to data come from

Loss of data authenticity and integrity (things at risk)

- Audiovisual content
- Descriptive Metadata
- Contracts, agreements, audit trail

© IT Innovation Centre and other partners of the FP7-ICT-2007-3 231161 PrestoPRIME consortium

The things that are at risk – and it's not just content, but also metadata and associated documentation

Loss of data authenticity and integrity (consequences)

- Loss of reputation
- Financial penalties (service provider)
- Extra time and resources needed to fix it again
- Loss of ability to use content (customer)
- Failure to record details of transactions with consequent denial by customer or service provider that they have agreed obligations

© IT Innovation Centre and other partners of the FP7-ICT-2007-3 231161 PrestoPRIME consortium

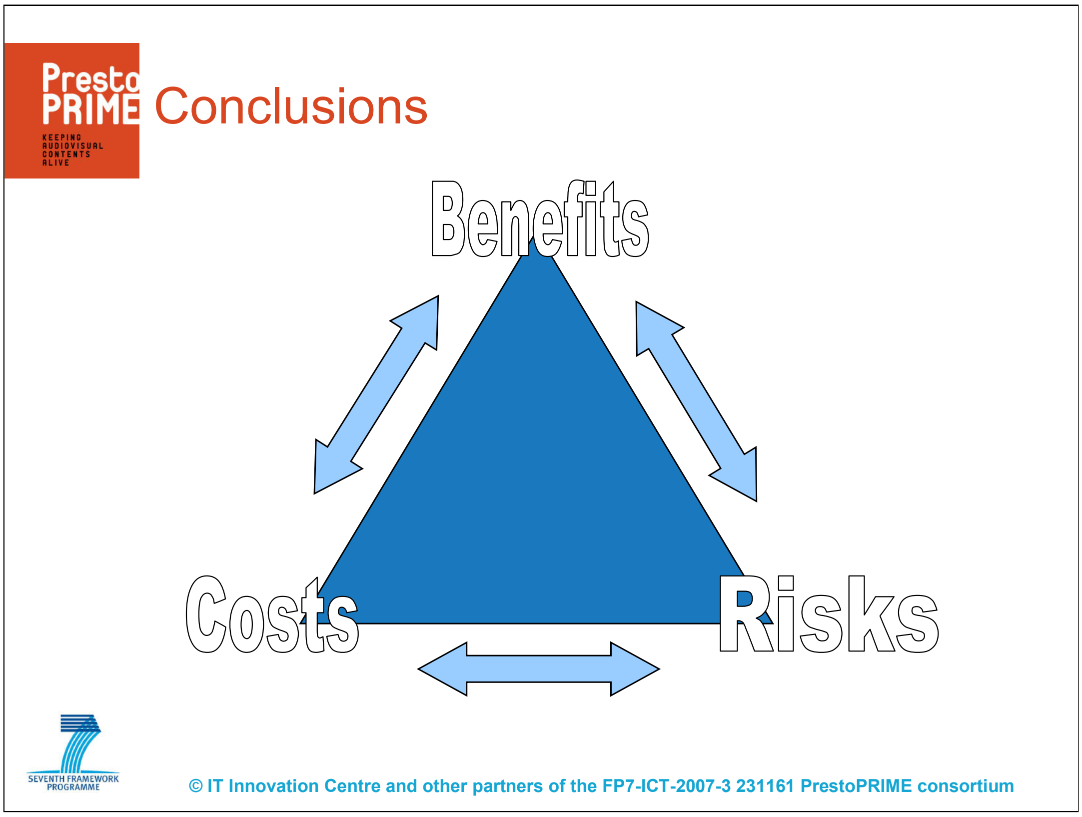
The consequences of the risk materialising, which isn't just data corruption or loss, but can include a much wider range of areas that all threaten archive operations

Loss of data authenticity and integrity (counter measures)

- **Enforce authentication and access control** so only trusted individuals have ability to manipulate assets (both within and external to the organisation)
- **Record all actions to content** that take place (who did what and when) to create a complete audit trail
- **Digital signatures (e.g. hashing) and integrity monitoring** to detect changes in digital content, both within storage systems and in transit over networks
- **Log any attempted breaches**, deliberate or accidental, and whether they were successful or not to allow security effectiveness to be measured.
- **Regular security audits** of technology, processes, staff skills etc.
- Evaluate and take into account any increased **risk from using data encryption** in storage systems as a potential degradation amplifier.
- Use appropriate **integrity assurance processes** that match the frequency, timescales and severity of the ways in which integrity could be lost
- **Ensure integrity records** (e.g. checksums or signatures) are **kept safe** and are themselves subject to integrity control
- **Ensure integrity control is comprehensive and consistent**, i.e. applied to all forms of data (metadata, identifiers, checksums, logs, credentials, audiovisual content)

© IT Innovation Centre and other partners of the FP7-ICT-2007-3 231161 PrestoPRIME consortium

And then finally some of the things that can be done to mitigate the risks.



OK, so now we're at the end. I started looking at some of the detailed issues of using IT storage technology applied to maintaining data integrity of audiovisual content.

There are clearly problems. Being aware of these problems and how to address them is a necessary first step.

Looking at hard drives, data tape and other storage approaches is a natural place to start as this is the closest tangible equivalent to video tapes, reels of film, etc. in many current archives. I've left out detailed recommendations on one technology v.s. another because this is all in the PrestoPRIME report.

How serious an issue this is will also vary from one archive to another, with some no doubt happy to accept that 'bit rot' might cause occasional frame loss in a big video sequence, but with others considering this an unthinkable outcome.

Then there are the wider body of risks to consider and the need to take a structured approach, i.e. risk management.

But whatever the level concerned, it still comes down to how much it costs, what is the risk of loss of content, and what is the benefit of incurring more cost to reduce this risk of loss.

Thank You and More Information

- http://www.prestoprime.org/
- Many public deliverables coming over the next month, some already on the website

- Scenarios, strategies, rights risks, migration, multivalent, storage, SLAs, workflows ….

- D2.1.1 Preservation Strategies
- D3.2.1 Threats to data integrity from use of large-scale data management environments

© IT Innovation Centre and other partners of the FP7-ICT-2007-3 231161 PrestoPRIME consortium

There's lots more information on the PrestoPRIME website which goes into a lot more detail on the areas I've mentioned. Some of the reports are already online and the rest should be available in a few weeks.