ICT-2007-3-231161

Deliverable D7.1.4

Audiovisual Digital Preservation Status Report 2

Richard Wright, BBC

15/02/2011

Document administrative table

Document Identifier	PP_WP7_D7_D7.1.4_Annual_AV_Status_2_R0	Release	0
Filename	PP_WP7_D7.1.4_Annual_AV_Status_2_R0_v1.00.pdf		
Workpackage and Task(s)	WP7 Dissemination and training T1 – Dissemination and publication of results		
Authors (company)	Richard Wright BBC		
Contributors (company)			
Internal Reviewers (company)	Jacqui Gupta (BBC); Beth Delaney (independent)		
Date	15/02/2011		
Status	Release		
Type	Deliverable		
Deliverable Nature	R = Report		
Dissemination Level	PU = Public		
Planned Deliv. Date	31/12/2010		
Actual Deliv. Date	15/02/2011		
Abstract	The current status of audiovisual preservation as of January 2011 is described. The 2009 report introduced the new problem of digital preservation (arising from the results of digitisation) and summarise the access issues for file-based audiovisual content and contributions of PrestoPRIME. This 2010 report concentrates on practicalities of audiovisual digital preservation: century costs, a digital preservation primer, and a summary of PrestoPRIME technology.		

DOCUMENT HISTORY

Release	Date	Reason of change	Status	Distribution
v0.01	10/12/2010	First Draft	incomplete	Confidential
v0.02	24/12/2010	Second Draft	BBC draft complete	Confidential
v0.03	28/12/2010	Third Draft	Circulated for partner contributions	Confidential
v0.04	06/02/2011	Fourth Draft	Complete – for review	Confidential
v0.05	14/02/2011	Final	Complete re-edit following review	Confidential
v1.00	15/02/2011	Completed	Release	Public

Table of Contents

Scope

PrestoPRIME is the European publicly-supported project that addresses **preservation of digital audiovisual content**, and **access to audiovisual content in digital libraries,** using **Europeana** as our demonstration platform.

This document is the sixth in a series of annual reviews of the status of audiovisual preservation in Europe. The first four reviews were produced by PrestoSpace. Each has had a specific focus, plus providing a general summary of annual progress toward saving Europe's audiovisual heritage.

The fifth was the first produced under the PrestoPRIME project, and covered:

1) introduction to PrestoPRIME;
2) Digitisation (of content not already in files);
3) Digital management and preservation: the problems of files;
4) Access.

This report deals with file-based content, and with practical issues in keeping audiovisual files usable for decades or centuries. The content of this report is summarised in the next section.

Executive summary

This document is a product of the EU-sponsored PrestoPRIME[1] project. PrestoPRIME is the major project on digital preservation in the audiovisual sector[2]. The current status of audiovisual preservation as of January 2011 is described. It is an update to the series of annual reports on audiovisual preservation previously given in January 2005 to 2008[3] as products of the EU-sponsored PrestoSpace project and it follows the PrestoPRIME report written in January 2010[4]. . The PrestoSpace reports concentrated on digitisation, which remains a significant issue. The January 2010 PrestoPRIME report surveyed activity in 2009 and introduced the new problem of digital preservation, which arises from the results of digitisation.

This January 2011 report (on the year 2010) has the following sections:

PrestoPRIME public activity: a summary of PrestoPRIME conference and workshop activity, including laying the foundations for PrestoCentre, the independent entity launching in March 2011 with the goal of providing *sustainable* support to audiovisual preservation.

Digital Preservation Primer: includes PrestoPRIME material presented at major archive conferences in 2010. The material concentrates on the 'trusted digital repository' concept, and so avoids getting lost in wondering whether broadcasters will use OAIS, or whether digital asset management systems (DAMs) will ever become digital preservation systems.

PrestoPRIME Tools: Descriptions of the tools produced by PrestoPRIME in 2010 and demonstrated at a public workshop in November.

The tools cover:

- Cost Estimation and Preservation Process Simulation
- Policy-based Storage Management
- Metadata Mapping and Validation services
- Video Quality Assessment
- Collecting and Integrating User-Generated Metadata
- Rights Ontology vs. Contracts
- Rosetta (including MXF Metadata Extractor)

Technology Updates: digital preservation technology doesn't stand still. Three areas have seen significant developments in 2010, as follows:

---

[1] http://www.prestoprime.org/

[2] PrestoPRIME is the only Integrated Project of audiovisual digital preservation running under the Seventh Framework of the EC-operated IST programme: http://cordis.europa.eu/fp7/ict/telearn-digicult/digicult-projects-prestoprime_en.html

[3] All four are online PDF files, available free from PrestoSpace. Three are listed here: http://digitalpreservation.ssl.co.uk/general/#White%20Paper , and the fourth is here: http://www.prestospace.org/project/deliverables/D22-9_Preservation_Status_2008

[4] The PrestoPRIME report from January 2010 is also free and online, listed here: https://prestoprimews.ina.fr/public/deliverables/PP_WP7_D7.1.3_Annual_AV_Status_R0_v1.00.pdf

- Containers – the level above files: how to handle groups of files. METS has been an accepted standard for Information Packages, but now there are several alternatives.
- LTFS – a filing system for data tape
- Digital-to-digital technology – audiovisual content has a particular category: digital but not in files. This data is on digital videotape, DAT, CD, DVD – and has preservation issues that have been the subject of new studies.

Cost of Digital Preservation: Century Store is a PrestoPRIME total-cost-of-ownership analysis of costs for keeping content for a century. The results show that digital storage for a century, including updating hardware and software and "continuous migration" of files, is not unaffordable, does not cost 11 times as much as 'writing back to film' – and indeed when access costs are considered, digital storage is very likely to be the cost-effective choice. This conclusion is important, because it is uncomfortable to have widely-quoted conclusions that differ both with common practice and common sense.

Digital Preservation of Broadcast Archives: relevance of PrestoPRIME: Audiovisual content is produced by many institutions (research, universities, the arts, medicine, defence, satellite imaging, undersea and geological exploration...) and held in many places (museums, galleries, libraries, archives, local history collections, oral history collections, government records offices, universities, national heritage institutions) in addition to broadcasting. However broadcasting does produce – and hold – about 40% of all audiovisual content.

Despite an expectation that PrestoPRIME and its predecessor projects would be biased toward the situation in broadcasting – the reality is:

- PrestoSpace concentrated on communicating the broadcast-based *preservation factory* approach (from Presto) to non-broadcast institutions;
- PrestoPRIME is about understanding the technology available from the digital library community (which certainly does *not* include broadcasting) and applying that to audiovisual content in general (not just broadcast content).

A major section of this Status Report addresses the real bias in the Presto series of projects: that it does *not* focus on broadcasting! We go through the existing digital preservation technology and the work of PrestoPRIME, and look at its relevance to, and potential impact for, broadcasting.

The *Relevance to Broadcasting* section covers:

- Where Digital Preservation Technology is Needed
- Fundamental Differences of Digital Content
- A Strategy for Digital Preservation
- Standards: OAIS, Europeana, Metadata
- Processes: Digitisation, Digital preservation, Management
- Tools
- Public Value
- Steps Toward Creating a Broadcast Digital Archive

1 A summary of PrestoPRIME public activity in 2010

This document is a status report on audiovisual preservation, not an advertisement for PrestoPRIME. The report concentrates on the situation of audiovisual content, and the technical needs and problems of all those who have responsibility for this content. However it is being produced by the PrestoPRIME project and the work of the project is meant to respond to these technical needs. So at various places in this document, mention will be made of relevant PrestoPRIME work.

The year 2010 was the second year of PrestoPRIME. The project moved from the specification stage (of 2009) to actual development of systems and tools for audiovisual preservation. Public information about these PrestoPRIME developments are given in the next three sections.

A major development of 2010 was the defining and announcement of the PrestoCentre the competence centre that will live on when PrestoPRIME ends. PrestoCentre is described in Section 1.4, below.

.1.1 Public Deliverables

The work of the first year (2009) of PrestoPRIME was formally reviewed in March 2010. Subsequent to that successful review, a set of documents was made public on the PrestoPRIME website http://ww.prestoprime.eu/project/public.en.html .

There are now (December 2010) 14 public deliverables, covering

- preservation requirements (D5.1.1 Definition of Scenarios),
- rights glossary,
- preservation status,
- preservation strategies,
- preservation process modelling,
- outsourced storage (D2.3.1 Service-Oriented Models for Audiovisual Content Storage),
- Europeana (D6.2.2 European Digital Library implementation guidelines for audiovisual archives),
- preservation toolkit
- four related major studies on preservation systems ID3.1.1 and threats ID3.2.1; use of emulation (Multivalent ID3.3.1) and services and service level agreement ID3.4.1.
- preservation technical architecture
- preservation metadata for audiovisual content

.1.2 Presentations

Issues within the general area of audiovisual digital preservation have been presented by PrestoPRIME partners at major conferences during 2010.

- Digital Preservation Interoperability Framework (DPIF) Symposium – 21-23 April 2010, Dresden, Germany; Gallo et al: 100 Million Hours of Audiovisual Content: Digital Preservation and Access in the PrestoPRIME Project

http://ddp.nist.gov/symposium/papers/02_10_Francesco_Gallo_PrestoPRIME_Digital_Preservation.pdf

- Joint Technical Symposium (JTS), http://www.jts2010.org/
  – 2-5 May 2010; Oslo, Norway.
  Matthew Addis: Long term data integrity for large Audiovisual archives
  Richard Wright (contributor): Migration of Media-Based Born-Digital Audiovisual Content to Files

- Transistor 2010 http://transistor.ciant.cz/2010/?page_id=120
  – 6-9 May 2010; Prague, Czech Republic
  Daniel Teruggi: PrestoSpace and Europeana
  Richard Wright: Digital preservation and PrestoPRIME

- IBC http://www.ibc.org/page.cfm/Link=46/t=m/goSection=5
  9-14 September 2010; Amsterdam, Holland
  Matthew Addis: Digital Preservation Strategies for AV Content
  (in a session on digital preservation chaired by Daniel Teruggi)

- FIAT/IFTA http://fiatifta-pilot.org/archives/1750
  15-18 October 2010; Dublin, Ireland
    - FIAT/IFTA pre-launch of PrestoCentre, the European Competence Centre for digitisation and digital preservation of AV content - by Jan Müller, Director of Netherlands Institute for Sound and Vision, the Netherlands. Included
      - Round Table: representatives of B&G, BBC, INA, ORF, RAI
      - PrestoCentre's online community by Marius Snyders, B&G
    - Preservation and Digitisation Session: clinic; Richard Wright, BBC
    - Testing Archive Systems; Christoph Bauer, ORF
    - PrestoPRIME poster: Daniel Teruggi, INA
    - Digitisation poster: Richard Wright, BBC
    - Producing and Archiving 3D poster: John Zubrzycki, BBC
    - PrestoPRIME workshop – see next section = workshops

- Symposium Vienna in Celebration of the 50th Anniversary of the Austrian Media Library (Österreichische Mediathek)
  27-28 October 2010; Vienna, Austria
  http://www.mediathek.at/ueber_die_mediathek/aktuelles_2.htm

- AVA21: Audiovisual Archives in the 21st Century Twitter: #AVA21
  13-14 October 2010; Ghent, Belgium
  http://www.ava21.be/en/ Chaired by Hans Westerhof, B&G; Presto, PrestoSpace and PrestoPRIME were the *only* audiovisual projects specificallly mentioned by the EC Commissioner Neelie Kroes.
  http://www.ava21.be/presentations/AVA21%20Executive%20Summary.pdf

http://www.ava21.be/presentations/AVA21%20Conference%20Report.pdf

- Arab States Broadcasting Union AUDIOVISUAL ARCHIVING SEMINAR participation by BBC, INA, B&G of PrestoPRIME
    - Philippe Poncin: Presto, PrestoSpace, PrestoPRIME
    - Johan Oomen: Metadata Enrichment
    - Richard Wright: Storing and Using Audiovisual Content

- IASA meeting jointly with AMIA http://www.iasa-conference.com
  http://www.amiaconference.com/2010/proposals-amia.htm
  2-6 November, Philadelphia, Pennsylvania USA
    - PrestoPRIME workshop – see next section = workshops
    - Richard Wright: Century Store
    - Richard Wright: A Collective Effort -- Online Audiovisual Resources in Europe

- INA/UCLA: Reimagining the Archive:
  12-14 November 2010; University of California, Los Angeles USA
  http://polaris.gseis.ucla.edu/reimagining/
  This conference was co-sponsored by INA, the coordinators of PrestoPRIME.
  Richard Wright: You Can Be Serious: Broadcast Archives and Academic Discourse

.1.3 Training and Workshops

During the Autumn of 2010, PrestoPRIME ran three one-day workshops.

- FIAT/IFTA 15-18 October 2010; Dublin, Ireland – a 3-hr workshop for approximately 30 delegates, covering:
    - digitisation
    - digital preservation
    - the ORF experience
    - PrestoPRIME competence centre

- IASA meeting jointly with AMIA 2-6 November, Philadelphia, PA USA
  a 4-hr workshop for 60 people, covering the FIAT content but also including presentations on:
    - Preserving Digital Public Television: Nan Rubin
    - Technical Architecture for Digital Preservation: Walter Allasia

  The FIAT/IFTA and IASA/AMIA workshop materials are online here:
  http://www.4shared.com/dir/TPEMV_n7/sharing.html

- PrestoPRIME Public Workshop 25-26 Nov London
  The previous public meeting was October 2009 in Vienna, reported here:
  http://www.prestoprime.eu/training/index.en.html

That meeting was largely about preservation needs, as identified during 2009 and as discussed at the workshop.

For 2010, PrestoPRIME has actual technology (described in more detail in Section 3, below). There were demonstrations of seven tools and systems, and general presentations which are now online, here: https://prestoprimews.ina.fr/public/presentations/london_26_11_2010 /

.1.4 Competence Centre

Projects come and go, but the problems of preserving audiovisual archives are here to stay. PrestoPRIME has the goal of establishing a source of information, cooperation, coordination and support that will be sustainable, and live long after PrestoPRIME. This entity is **PrestoCentre**, a competence centre created and launched by PrestoPRIME.

During 2010 the foundation was created , and the concept was announced at the FIAT-IFTA conference in Dublin in November. A full session was devoted to this announcement, with short speeches supporting the Centre from the five archives working together to create PrestoCentre

- British Broadcasting Corporation (BBC);
- l'Institut National de l'Audiovisuel (INA);
- Netherlands Institute for Sound and Vision (NISV;
- Österreichischer Rundfunk (ORF);
- Radiotelevisione Italiana (RAI).

For more than a decade, these archives have worked together in the EU-funded 'Presto' series of projects to bring together expertise and experience in AV digitisation and preservation in Europe. The goal of PrestoCentre is to make sure that the knowledge and dedication that was built up in these projects persists and does not 'fade away'. PrestoCentre is a membership driven, non-profit organisation that will serve stakeholders in audiovisual digitisation and digital preservation in Europe.

The launch was introduced by Jan Müller, head of the Netherlands Institute of Sound and Vision (NISV). That presentation[5] and the whole launch session are online[6]. A detailed introduction to PrestoCentre[7] was provided by Marius Snyders of NISV, who is leading the work in PrestoPRIME to launch the PrestoCentre.

PrestoPRIME is holding a large international conference **Screening the Future** in Hilversum, The Netherlands on 14-15 March 2011. The purpose is to shape the agenda of the Competence Centre and to bring as much attention as possible to the new service in Europe. The conference delegates will include small and large archives, service providers, vendors, funders, policymakers and educators developing solutions to the most urgent questions facing audiovisual archiving.

[5] http://fiatifta-pilot.org/archives/2343
[6] http://fiatifta-pilot.org/archives/2103
[7] http://fiatifta-pilot.org/archives/2355

2 A Digital Preservation Primer

> "Digital preservation requires the indefinite error-free storage of digital information, with means for its retrieval and interpretation, irrespective of changes in technologies, support and data formats, or changes in the requirements of the user community."[8]

PrestoSpace produced a great deal of guidance on conservation and digitisation, summarised in the PrestoSpace wiki: http://wiki.prestospace.org/ .

PrestoPRIME has a range of materials on digital preservation strategy and planning available as documents on the project website http://www.prestoprime.eu/project/public.en.html – and has just run two workshops on digital preservation.

The materials from the workshops are also available online: http://www.4shared.com/dir/TPEMV_n7/sharing.html

The problem with all the above is that it tries to be comprehensive, and so for some purposes – such as getting started in digital preservation – it's just too much. The following primer presents a broader, although less comprehensive overview . It does have references to more detailed information.

.2.1 Conservation

"Digitisation and transfer processes actual occupy a tiny proportion of the lifetime of an object. For the majority of the time, the main issue is conservation."[9]

In the file-based world the word curation is more common than conservation, but the task is the same: maintaining what you have.

There are four main factors in a programme of conservation[10]:

1. Handling, packaging and storing
2. Environmental conditions
3. Protecting the masters
4. Condition monitoring

I was going to write two sections: analogue conservation and digital conservation – but the above list (with appropriate adjustments) applies to both. Full details on the above four points, for shelf-based content, are all in the PrestoSpace wiki (being revised for the PrestoCentre Competence Centre). For digital content there are new considerations under the same headings:

---

[8] Consultative Committee for Space Data Systems. (2002). Reference Model for an Open Archival Information System (OAIS). Washington, DC: CCSDS Secretariat, p. 1-1

[9] http://wiki.prestospace.org/pmwiki.php?n=Main.PreservationStrategy#Conservation

[10] Author's note: I was going to write two sections: analogue conservation and digital conservation – but the above list (with appropriate adjustments) applies to both.

.2.1.1 Handling, packaging and storing

Handling: how do you move files around? Do you have a way to check that a file is the same after it has been moved? Use of a *trusted digital repository* (or some sort of *asset management system* may give you handling tools).

The simplest approach is a *fixity check,* and the simplest of those is the checksum. This approach requires:

- creating a kind of key from a file known to be good,
- keeping that key separate from the file itself, in a log or register,
- checking that key against the same key recreated from a copy or any other subsequent us of the file. If the keys match, the files match (to within a certain probability, which can be made as high as needed by increasing the complexity of the key) and so the file is good, and the *handling* (moving to new storage; making a copy for someone in another place, or even just reading the information back from storage) has been successful.

A fixity check is a well-defined concept in digital preservation. It is a *preservation event*. Calculating the fixity is an event that has to be noted (and guaranteed to happen, at the right time) – and checking the fixity is a related event, again something that has to be guaranteed to happen. Formal repositories, and in particular digital preservation systems, ensure that fixity calculation and checking does indeed take place at just those times and places necessary to ensure against getting incorrect versions of a file, or getting a corrupt file.

More information[11] about *fixity* as a formal concept is available from the Library of Congress and the other cited resources.

If properly implemented and used, a fixity check provides as much security as can be achieved regarding assurance that a file has not been changed, and provides a reliable flag to show when a file has been changed. What a fixity check *cannot* do is repair a file if bit-rot or a read-back problem or a data transmission problem or some other error has caused a change. Hence the need for more than one copy of all files in a preservation system, so that ideally the fixity calculated from another copy will match the value recorded in a log of fixity data, and so can be assumed to be intact.

Packaging: Audiovisual file-based content is packaged in *wrappers.* A wrapper is just a kind of file that is complicated enough to hold the various things that audiovisual content needs:

- the (digital) audio and video signals themselves; often multiple audio signals and sometimes multiple video signals (multiple clips or a stereo-optical signal as used for 3D)
- associated time-based data such as timecode
- associated time-based metadata such as subtitles (closed captions)
- metadata about the whole file, of various sorts

[11] Definition of Fixity Check: http://id.loc.gov/vocabulary/preservationEvents/fixityCheck.html Further details about implementation: http://digitalpreservation.ncdcr.gov/newtodp.html and http://archivematica.org/wiki/index.php?title=Overview

Part of *packaging* includes the way the audio and video are encoded from signals into numbers. Whatever encoding is chosen, will in turn affect how the file can be used in the future, because a decoder will be needed to play the file. A related issue is use of compression (data reduction). Dramatic reductions in file size can be achieved by throwing away parts of the signal that are near-repeats of other parts of the signal, but quality is eventually lost. Further, a data-reduced signal is less robust and so will then be less useful in any operation (such as an edit) that involves another cycle of decode-encode. A compressed signal is also more affected by small errors in the file.

The choice of encoding type and wrapper format rapidly become complicated issues.

There are now three main kinds of wrapper (and many many more less frequently used wrappers):

- The MAC world, which uses MOV;
- The PC world, which uses AVI;
- Certain professional areas (broadcasting, digital cinema) which use the non-proprietary SMPTE-standard wrapper, MXF

All are capable of holding various kinds of encodings, lossy and lossless. All have some degree of interoperability (meaning MACs can make and play AVI, and PCs can make and play MOV, and both can make MXF).

PrestoPRIME recommends MXF, and is working to make it simpler to use and with a greater range of supporting tools. However both MOV and AVI are in use by leading audiovisual archives.

Which kind of coding? PrestoPRIME argues (Section 6.3.2) that the endgame for audiovisual preservation is uncompressed signals, rather than use of lossless or lossy compression. We also supply a roadmap[12] for getting to uncompressed, and have detailed flowcharts[13] for the decision-making needed while on that roadmap.

The essential strategy is:

- never go down in quality
- never go sideways in quality = transcode, because in practice that will be a generation loss which amounts to 'going down'
- never get boxed in (by proprietary and obsolete technology)

Students of thermodynamics will recognise these statements as re-phrasings of the first three laws of thermodynamics as understood by generations of undergraduates:

- you can't win
- you can't break even
- you can't get out of the game

[12] http://wiki.prestospace.org/pmwiki.php?n=Main.Roadmap
[13] https://prestoprimews.ina.fr/public/deliverables/PP_WP2_D2.1.1_preservationstrategies_R0_v1.00.pdf

Storing: PrestoSpace provided a lot of information on storage[14], and so has IASA in its TC-04 publication[15]. Again, there is so much there that the very amount of information causes a problem.

Very basic guidance on storage:

- Use mass storage, not CD / DVD / BlueRay.
- *The Cloud*[16] remains expensive (2 to 5x more than local storage)
- Choice of two: hard drives, tape

Tape or hard drives:

- BOTH need migration every 5 years
- BOTH have failures
- TAPE has lower failure rate than hard drives
- TAPE has higher "start-up costs"

Conclusion: use hard drives until you have enough data to 'amortise' a tape drive ($2000 or so)

Finally, the guidance for decades for secure storage has been[17]:

- two copies
- on two technologies
- in two places

There is now growing use of outsourced storage, which should be good news to archivists who do not want to become digital storage experts. A major proposition for taking care of audiovisual content is AVAN[18], and the PrestoCentre will support the concept of share-and-conquer as one solution to storage. PrestoPRIME has technical information and tools[19] about actual detailed management of contracted storage. Also see Section for more on the storage management tools.

.2.1.2 Environmental conditions

In the real world, this means temperature and humidity control, fire prevention and a good roof. In the strange new world of invisible archives on mass storage, there is the issue of the *social, political and economic* environment – the funding and management of this invisible archive.

---

[14] http://digitalpreservation.ssl.co.uk/
[15] http://www.iasa-web.org/tc04/archival-storage
[16] Cloud Storage http://en.wikipedia.org/wiki/Cloud_storage
[17] http://www.docstoc.com/docs/2567237/How-to-Preserve-Audio-(and-video).This from the author, but it seems to have propagated across the web, so somebody must think it's worth looking at; see slide 15.
[18] http://www.archivenetwork.org/
[19] https://prestoprimews.ina.fr/public/deliverables/PP_WP2_D2.3.1_SOAforAV_R1_v1.01.pdf

Digital libraries have developed the concept of a *trusted digital repository*, and have gone further to develop check-lists of what needs to be in place to gain the trust of users of such a repository.

The standard check-list is TRAC[20], and here are the main points:

TRAC Criteria Documents

A1.2	Contingency plans, succession plans, escrow arrangements (as appropriate)
A3.1	Definition of designated community(ies), and policy relating to service levels
A3.3	Policies relating to legal permissions
A3.5	Policies and procedures relating to feedback
A4.3	Financial procedures
A5.5	Policies/procedures relating to challenges to rights (only if likely to be needed)
B1	Procedures related to ingest
B2.10	Process for testing understandability
B4.1	Preservation strategies
B4.2	Storage/migration strategies
B6.2	Policy for recording access actions
B6.4	Policy for access
C1.7	Processes for media change
C1.8	Change management process
C1.9	Critical change test process
C1.10	Security update process
C2.1	Process to monitor required changes to hardware
C2.2	Process to monitor required changes to software
C3.4	Disaster plans

.2.1.3 Protecting the masters

This was a major issue in shelf-based archives. Proxies or viewing/listening copies were essential in order to minimise use of precious and fragile masters. Now we can make as many master copies as we can afford to store, but we may find that NONE of them are suitable for access purposes.

In consequence, proxies are still an issue in digital archives, but for new reasons. The master file for video will be very large: impossibly large for web access, and probably too large for efficient access even within the same building (as the storage) over existing data networks.

Digital audiovisual archives need to have *viewing proxies,* usually of two sorts:

- medium quality for in-house professional use
- low quality for web access

One problem is that bandwidth (network capacity within an institution via Ethernet; network capacity to people's home via Internet) keeps increasing, and so the requirements for these proxies change over just a few years. Also the preferred encodings change: the BBC started with Real Video, moved to some use of MPEG-2 and Windows Media, and now is using Flash Video.

[20] TRAC information: http://en.wikipedia.org/wiki/Trustworthy_Repositories_Audit_%26_Certification
TRAC definition document:
http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf

The changes in requirements for proxies causes archives to have to repeatedly “go back to the masters” to make new proxies. This process doesn't involve risk of damage to the masters as it would have in analogue days, but it does involve time and expense. An efficient approach is to have three levels:

- master;
- mezzanine: the most efficient coding for generating new proxies;
- viewing proxies (which could be of various quality levels, so really there are a lot of levels in a fully-developed implementation).

For instance, the EDCine project[21] uses lossless JPEG2000 (in an MXF wrapper) for the master, and a high-quality lossy JPEG2000 encoding for the mezzanine, because lossy JPEG2000 supports very efficient coding of the lower-quality proxies – in particular the distribution format required for digital cinema.

.2.1.4 Condition monitoring

In the analogue world, there was a need to 'check the shelves' do see that stock was all present and accounted for, and in good conditions – and there was technology such as a A-D Strips[22] that could monitor for acetic acid.

In our new world, storage media (which have many properties in common with analogue audio and video media) also deteriorates, and is subject to various other sources of failure. However the greatest risks are:

- the overall complexity of computer-based systems
- the rapid obsolescence at all levels of these systems

If storage is successfully outsourced, condition monitoring is someone else's problem – ideally an IT professional that will control risks more effectively than would professional archivists who may also be quite amateur storage managers.

Either way, within professional storage management there are approaches for monitoring *what's happening* at various levels, including:

- technology built into storage media hardware[23] for measuring errors and anticipating (and hopefully preventing) media failure;
- storage management software that monitors performance and errors, again anticipating problems in order to prevent failures[24];
- processes within storage management software that periodically test content to check that it is readable and correct (*scrubbing*)

All this costs money, and possibly time as well. For in-house storage, somebody in charge needs to understand at least the three layers of condition monitoring just mentioned, or archive content could be needlessly lost. For outsourced storage, it is useful to understand enough about the above in order to understand why managed

[21] http://www.edcine.org/documents/public/edcine_tcf_ds_d1_2_v3_ucl.pdf/view
[22] http://www.filmpreservation.org/preservation-basics/vinegar-syndrome
[23] E.g. SMART http://en.wikipedia.org/wiki/S.M.A.R.T.
[24] For instance, StorSentry: http://www.hi-stor.com/site/ This mention is not an endorsement

storage costs what it does, and what is gained by investment in these various technologies.

PrestoPRIME's storage management tool (see Section) shows the beneficial results of use of *scrubbing*, at various time intervals.

.2.2 Digitisation

There are two questions: why, and how. PrestoSpace published "Why Digitise?[25]" as an online tutorial. A main reason is the obsolescence of analogue audio and video carriers, so film is always a special case,

PrestoPRIME has presented a simple four-point guide to how to digitise:

- Save the original
- Digitise @ SDI = 4:2:2 = 200 megabits/sec
- Save exactly as digitised = uncompressed
- Use an open source file format (MXF or ?)

The following sections expand on these points.

.2.2.1 Save the original

The recommendation is to *always* save the original, despite all the following problems:

- Cost of storage (especially cooling)
- Obsolescence of players (especially for video)
- Lack of a business case
- Low probability of ever being used

Short-term reasons for saving the original:

- Provides a back-up position for errors or problems in the migration / digitisation
- Provides a quality reference
- "Errors" are easy, compared to maintaining quality, which isn't even measured for most audio/video transfers
- Provides an additional copy

Long-term reasons for saving the original:

- All the short-term ones
- Supports 'evidence-based destruction' of the originals, only AFTER years (or decades) of experience with the migrated 'new master'

.2.2.2 Digitise @ SDI = 4:2:2 = 200 megabits/sec

SDI[26] means Serial Digital Interface, the broadcast industry standard for digital video, that was first written over 25 years ago (so it's stable, well-known and well-supported).

---

[25] http://digitalpreservation.ssl.co.uk/general/T6/T6-1.html
[26] http://en.wikipedia.org/wiki/Serial_digital_interface

The recommendation to use SDI as the encoding standard has two parts:

- it's only recommended for standard-definition video, meaning television as we knew it up until the current decade, but NOT high-definition video that is now rapidly replacing standard-definition. Most audiovisual archives consist mainly of SD video, because HD has only recently begun to arrive at the archive.
- SDI is uncompressed. As previously stated, the *end game* for audiovisual preservation is uncompressed signals, and use of SDI gets us straight there.

The needed technology is a computer-based *capture card* that accepts the output provided by the player (of the videotape being digitised). Immediately there are complexities that take this discussion beyond the bounds of a *primer:*

- Component vs. composite: the analogue tape may be compositem, and so will need decoding into component form for best digitisation.
- Correction: the signal from the source can be impaired in various ways, and should be corrected using various professional tools: a time-base corrector, possibly a time-code regenerator, possibly one of more filters (of unwanted noise that will affect the quality of the digitisation)

The target standard – SDI – is easy enough to put into a primer, but the rest of video digitisation is a professional task for persons who really are professionals!

The BBC has open-source software for capture of and SDI signal and putting it into a very simple form of MXF wrapper: the INGEX[27] project.

.2.2.3 Save exactly as digitised = uncompressed

Having digitised and created an uncompressed signal, the obvious thing to do is to save all those bits, just as they are. However a factor of approximately three in storage requirements can be saved by use of lossless encoding using JPEG2000, and this is an approach that has been adopted by the Library of Congress and others.

The decision about uncompressed vs. lossless-compressed should be based on all the facts. It is a decision about economy and workflow, and so the costs of implementing the JPEG2000 processing (on input, and on every subsequent use of the master) needs to be included, as a factor weighed against the savings in storage.

One issue to remember is that the 'overhead' of using JPEG2000 will remain as a continuing processing-time overhead, and replacement/maintenance of hardware encoders and licencing for software encoders will also remain as items costing as much or more in the future as they do now, while storage will continue to drop in cost for the foreseeable future (at least the next 20 years).

For low-quality originals, such as VHS, the PrestoSpace roadmap (Footnote 12) suggests encoding at DV quality (25 megabits/second). This is a ‘mezzanine’ approach, because DV is a high-quality lossy encoding that is probably the most widely-supported lossy video format in existence.

[27] http://ingex.sourceforge.net/archive/

.2.2.4 Use an open standard file format: (MXF or what?)

There are really three considerations

- Open standard
- Well-supported
- Well-documented

Use of a file format based on an open standard is important for keeping independent of proprietary products, for not being 'boxed in'. However there is some conflict between current formats which are open standard (e.g. MXF) and the desire to also use a format that is well-supported and well-documented. The Apple QuickTime MOV format is not officially an open standard, though there is a relationship between QuickTime and MPEG-4. Further, both Apple and Microsoft publish a lot of detail about the MOV (Apple) and AVI (Microsoft) formats.

It should be remembered that WAV is also a Microsoft format, which has not been any impediment to its use as the basis for the EBU and AES standardisation of the Broadcast Wave Format[28].

Reasons to consider MXF:

- a SMPTE standard
- supported by the EBU
- widely used in broadcasting: D10, IMX
- the standard wrapper for digital cinema
- technical support from a range of open-source and commercial products: MOG, Opencube, MFXLib
- Supported by standard editor software: Adobe Premiere Pro, Apple Final Cut Pro, Avid Media Composer ...
- Used in several major video preservation projects: Library of Congress, BBC

The BBC INGEX[29] software provides open-source support for MXF OP-1A, the simplest version (profile) of the use of MXF and so it is recommended for archiving (where simplicity is always a virtue). The Library of Congress is coordinating the future development of MXF for the purposes of the US federal government through the Federal Agencies Digitization Guidelines Initiative (FADGI)[30]. FADGI is working with the group that has been responsible for MXF and its predecessor (and big brother) AAF, the Advanced Media Workflow Association[31] (AMWA). AMWA will work on development of specific application and application profiles for MXF. The purpose of the work is to ensure clear, unambiguous, fully-compatible MXF implementations. Such clarity is essential for long-term preservation – the last thing anyone needs in an archive is a file retrieved years or decades later, and found to be the ‘wrong kind of MXF’.

[28] http://www.ebu.ch/fr/technical/publications/userguides/bwf_user_guide.php
[29] http://ingex.sourceforge.net/archive/
[30] http://www.digitizationguidelines.gov
[31] http://www.amwa.tv/

3 PrestoPRIME Technology

A public workshop was held in November where a range of audiovisual preservation tools was demonstrated to a capacity audience of about sixty persons (the largest meeting ever in the new BBC R&D premises). There was a group session with explanatory talks, and then live demonstrations in various rooms (and corners of rooms so delegates could circulate at will). The slides from the talks are online, as are additional materials from some of the demonstrations, as follows:

And introductory talk covering all the technology and the overall plan of work in PrestoPRIME was given by Laurent Boch (RAI), "How do AV files really get saved?" PrestoPRIME_20101126_1_HowDoAVFilesReallyGetSaved.pdf

The specific tools and demonstrations were:

Cost Estimation Tools – online tools supporting decisions on storage planning, file migration, coding decisions – and their consequences.

Demonstration by Matthew Addis and Mariusz Jacyno, IT Innovation
PrestoPRIME_20101126_2_WP2_StrategyPlanningTools.pdf
Presentation by Richard Wright (BBC), "Preservation strategy and planning tools".

Policy-based storage management – we speak of 'storage as a service' but what does it really mean, and how are services managed?

PrestoPRIME_20101126_3_WP3_ControlOfStorageServicesQA.pdf
Demonstration and Presentation by Stephen Phillips (IT Innovation), "Control of storage and services, quality analysis".

Metadata Mapping and Validation services – demonstrating mapping and validation of metadata from multiple standards**.**

PrestoPRIME_20101126_4_WP4_MetadataContentTrackingRights.pdf
Demonstration and Presentation by Werner Bailer (Joanneum Research), "Metadata handling: mapping, rights, provenance".

Video Quality Assessment – a fully-automated approach to identifying a broad range of defects in video

PrestoPRIME_20101126_Demo_VideoQuality.pdf
Demonstration by Peter Schallauer (Joanneum Research), "Video Quality Assessment".

Collecting and integrating user-generated metadata – how do we use 'the crowd'? Collecting and curating crowd-sourced metadata

PrestoPRIME_20101126_Demo_UserGeneratedMetadata.pdf
Demonstration by Michiel Hildebrand (VUA) : "Collecting and integrating user-generated metadata".

Rights Ontology vs. Contracts – having people interpret contracts and clear rights doesn't scale (to the problem of accessing million-hour broadcast archives). This tool

is a formal language for unambiguous determination of rights, supporting automation of rights clearance.

PrestoPRIME_20101126_Demo_RightsOntology.pdf
Demonstration by Laurent Boch (RAI) and Francesco Gallo (Eurix), "Rights Ontology vs. Contracts".

Rosetta (including MXF metadata extractor) – Rosetta is a commercial system for digital preservation (one of about two in the world). The demonstration showed how it works, and how it copes with broadcast formats such as MXF.

PrestoPRIME_20101126_5_WP5.pdf
Presentation by Francesco Gallo (Eurix) and Nir kashi (ExLibris), "Organisation of digital preservation framework".

Beyond tools:
Do we need anything special to preserve files? That question was addressed in the presentation "**A reality check** - Standard IT processes that have been in use for decades" by Matthew Addis (IT Innovation)
PrestoPRIME_20101126_RealityCheck_1_StandardITProcesses.pdf
The answer was: we don't need anything special:

- If we only have a small number of files
- If we don't care how much it costs to maintain digital objects, and so are willing to pay for unlimited amounts of manual intervention
- If we have our own skilled IT specialists who can integrate all the existing tools, processes and systems, make them work and keep them working

For the rest of us, digital preservation technology remains interesting.

The day concluded with a presentation on the **PrestoCentre Competence Centre**
PrestoPRIME_20101126_PrestoCentrePresentation.pdf
Presentation by Marius Snyders (Beeld en Geluid), "the PrestoCentre : Competence Centre on Digital Preservation and Migration".

The PrestoCentre has a major role to play regarding tools: making sure they work, documenting and explaining them, licensing them where appropriate, and seeing to it that the tools are maintained and updated for so long as they have value. Digital tools need digital preservation, too!

4 Technology – Brief Encounters

This section gives short summaries of recent developments in areas where audiovisual preservation technology and practice is rapidly developing. The next three sections cover:

1) Preservation of groups of related objects: containers for multiple files;
2) A file system for datatape;
3) Moving digital data from non-file-based to file-based storage.

An interesting point about these topics is the overlap: files are held in filing systems and groups of files (topic one) can be held (on disc) in folders or other structures that use filing systems. But groups of files are held on datatape in proprietary systems that are very different from standard computer filing systems. Optionally, files or groups of files can be written to datatape as a UNIX (open standard) TAR object: a TARball, which still isn't a standard computer filing system.

Hence the importance of a standard filing system for datatape, topic two. But there are other examples of digital data not in filing systems, which is a particular issue for videotape collections as digital video tape (e.g. DV, Digibeta) will need to be migrated to file-based storage (topic three).

.4.1 Containers

We talk about audiovisual content moving from tapes and reels into files, and in digital preservation we worry about the continued existence and usability of files. However the collections of information moving into digital archives commonly now consist of multiple, related files. This fact introduces a new level for organisation of information: the **container**, meaning any formal way to deal with multiple objects at once, which in a file-based world means dealing with files in groups.

The issue arises in digital archives and libraries, because as we build collections we immediately find that either:

- A single object has multiple kinds of information that could be in multiple files (such as sound and video in separate files) – or:
- The basic unit in the collection starts off as multiple objects: a television programme can be contracts, script (various versions), production documentation and correspondence (lots), publicity stills and a range of different broadcast versions (UK and US formats, subtitles or dubs in various languages, re-edit for rebroadcast with longer or shorter duration and on and on) plus any number of other related objects that are all part of the 'television programme'.

The simplest unit above the level of the file is a folder. There is audiovisual content that uses folders: the DPX[32] format for film uses one file for each frame, and the frames for a clip or entire production are held in a folder. Unfortunately there is no technology to control the folder level: nothing (except constraints in the applications reading from and writing to the folder) to control that all the frames are there, and in the right order.

[32] DPX http://www.cineon.com/ff_draft.php#tv http://netghost.narod.ru/gff/graphics/summary/dpx.htm

The Library of Congress and their digital library partners came up with a simple and elegant solution to the problem of contribution of multiple related files as a single digital library unit: the BagIt[33] tools for transferring electronic 'bags' of files. While designed for getting related objects from one institution to another, they also provide a simple container for any other purpose (such as a unit for storage).

BagIt includes important information that a folder does not have:

- An inventory of contents, and built-in inventory checking (solving the 'all present and accounted for' problem);
- A structure designed to be system independent (so PC and MAC and UNIX users can all share a common bag);
- Machine readable metadata, to assist and secure all transmission (or storage) processes applied to the bag.

The container issue is very clear in digital preservation technology, because the OAIS standard is based on Information Packages, and almost by definition such a package should at least support multiple files. The METS standard has been widely adopted for Information Packages in systems following OAIS – though not universally. The Stanford Digital Repository has dropped METS in its recent revision[34], citing the high overhead required to create METS SIPS at input and to change from a METS AIP to a wide range of METS DIPs at output. In their new system, METS is confined to use as one of a range of supported DIP formats. Instead of METS, they are using BagIt and the SIP, and embedding BagIt support right down at the basic data model level[35].

Another simple alternative to METS, which has been developed specifically for audiovisual purposes by the Japanese national broadcaster NHK, has now been standardised by ARIB[36], the professional association for the broadcasting industry in (mainly) Japan. Their proposal — ARIB TR-B31 — is now being considered for SMPTE standardisation. It has been developed with full awareness of the BagIt work, the strengths and weaknesses of METS, and the need for use of files and open standards while still supporting tape-based transfers.

Accordingly, a key element of TR-B31 is availability of a way to put *organised groups of files* (the BagIt concept) onto datatape – *as files, in an open standard*. The last requirement needed a new way to use datatape, which is presented in the next section on LTFS.

TR-B31 defines (in XML) structures needed in broadcasting. The basic unit of *programme* has the following elements[37]:

- Titles: a program title

---

[33] http://www.digitalpreservation.gov/videos/bagit0609.html Bagit: Transferring Content for Digital Preservation

[34] Tom Cramer, Stanford University Libraries: "Designing and Implementing Second Generation Digital Preservation Services: A Scalable Model for the Stanford Digital Repository" D-Lib Magazine, Sept/Oct 2010, Vol 16, No 9/10 http://www.dlib.org/dlib/september10/cramer/09cramer.html

[35] Cramer op cit, Fig 5

[36] http://www.arib.or.jp/english/

[37] Association of Radio Industries and Businesses (ARIB) Technical Report TR-B31, Section 6. http://www.arib.or.jp/english/html/overview/doc/4-TR-B31v1_0.pdf

- Identification: an identification for a program or a file
- GroupRelationship: a description of relationship between programs or files
- Event: a start and end date/time of program on-air and such.
- Publication: a description about a media or a service for an event and such
- Annotation: an annotation of a program
- Classification: a description for video or audio material classification
- Participant: a description for an associated person or organization
- Person: a description for a specific person
- Organization: a description for an organization
- Address: an address associated with a person or an organization
- Communications: a method for communication such as telephone number.
- Contract: a description for an contact of a program or material
- Rights: a description for copyrights of a video/audio material
- Playlist: a sequence of a program that is delivered with multiple roles.
- VideoDescription: a description for a video essence element
- AudioDescription: a description for an audio essence element
- CaptionsDescription: a description for a closed caption element
- AncillaryDescription: a description for an ancillary data element
- FileDescription: a description for a file element
- Block: a description for a successive block in a video or audio essence
- Keypoint: a description for a specific part in an essence
- CueSheet: a description for cue-sheet

There are two approaches for further development of complex objects or groups of files as one unit: wrappers and containers. The audiovisual world already knows how to get multiple kinds of information into one well-organised and secure place: use a wrapper. A wrapper provides a tight way to hold information, as compared to a folder which is just a bare receptacle with no functionality or security built-in. It takes work to make the sound disappear from a MOV or AVI or MXF file that holds video and audio. It just takes one press of the delete key to remove the sound file from a folder holding audio and video files.

Now the development of formal containers (METS, BagIt, TR-B31) gives the audiovisual world a new option. We can continue to put more and more into wrapper files, or we can move up from the file level to the container level – where there is more flexibility while still having needed security and efficiency,

Work is continuing on the MXF format, to make its specification tighter through development of application profiles (LOC and AMWA work, previously described in Section 2.2.4). There is every reason to welcome such work. However there is also the temptation, whenever a standard is being reviewed, to also add in new functionality. After all, anything could go into an MXF file, including the whole list of TR-B31 Programme entities listed above. **But where should additional complexity reside?**

In principle, the PrestoPRIME view is that MXF should be simplified so that it is easier to use and therefore will have wider use. The LOC/AMWA work to define Application Profiles can be seen as simplification: eliminating sources of ambiguity and confusion, and therefore making use of MXF simpler.

But for all the other things that audiovisual collections want to put into secure storage and into a digital library or trusted repository – as a unit – the use of containers is clearer, cleaner and simpler than the alternative of putting more information, and more kinds of information, into wrapper files.

Here is a taxonomy of digital objects:

Table 1 - Kinds of digital objects

Objects	file-based	NOT file-based
individual object	file, wrapper	DV, DAT tape
group of objects	Bagit, METS, ARIB (on disc, or using LTFS on datatape)	TARball on datatape; audio CD

We want to move data from the right-hand column into the centre column, and how to do that is the subject of Section 4.3, below. We also want to use datatape, because it is still the cheapest way to store data, and audiovisual collections have lots of data. Unfortunately datatape was not a proper file-based medium until the advent of the LTFS system, explained in the next section.

But in moving this non-file data, and in creating *trusted digital objects* for preservation, we still have the choice between two different rows in the table: using wrappers or using containers.

There are many reasons to consider containers as the preferred choice for future development, and to abandon any further complexity for wrappers – they're complicated enough already!

.4.2 LTFS – putting files on datatape

If you were to buy a datatape drive and a stack of datatapes and plug the drive into your computer, chances are that the drive would remain invisible to you and to the rest of the computer system. Datatape was different: it didn't support random access, so it had its own storage management rather than being seen as fully-compatible with file-based storage systems.

It is a bit puzzling how this situation arose, because although the original reels of 7 and 9-track IBM datatape from the 1950's and 1960's were certainly sequential storage media, there was use of datatape in the 1960's with pre-formatted blocks of fixed size and location, allowing blocks of data on such a tape reel to be entirely compatible with, for instance, a similar set of blocks on a floppy disc or hard drive. The pioneer of this approach was LINC-tape, developed at MIT Lincoln Labs in the early 1960's, and made more widely available by the Digital Equipment Corporation as DEC-tape (available on DEC computers, including the very odd PDP-12 computer, which could convert from running as a DEC machine to running as a LINC machine)[38].

Somehow by modern times datatape had lost all compatibility with other forms of storage, and required special commands or applications to 'drive the tape unit'. The

[38] http://en.wikipedia.org/wiki/LINC

result was a given type of datatape system (tape drive or library) usually required a commitment to a proprietary method of reading and writing the tape.

This fact seriously compromised the usefulness of tape in two areas:

- Exchange: tape could only be exchanged between identical library systems, and even then it was hard to know where content in a tape library would physically reside. What was to the user a single file could be split across multiple datatapes;
- Preservation: a basic principle for keeping data long-term is use of open standards and avoidance of proprietary systems.

A partial solution was to use the UNIX TAR format (as in the BBC INGEX approach to transferring video from videotape to datatape), but this wasn't much of a solution to the non-UNIX world.

Now there is an open standard – LTFS[39] – for file-based use of LTO-5 datatape, supported by several companies (IBM, HP, Quantum) and by several tape manufacturers (Tandberg Data, Fujifilm, Imation, Maxell, Sony, TDK).

An LTFS drive plugged into a computer will look like, and act like, any other external storage unit (e.g. flash drive, external hard drive). Files can be written to it, and groups of files supporting the kind of data that we need to keep together as a 'complex digital object' can be written to the file using a container format (such as BagIt or the TR-B31 standard described in the previous section).

The tape in the drive can then be sent to anyone with an LTFS tape drive and the data can be read back, with no additional complexity and no dependency on tape library software or indeed on any additional software, proprietary or otherwise. Datatape has become easily and generally useful – and I can now stop buying newer and bigger hard drives every few months, and start to seriously consider datatape even for home use.

.4.3 Digital-to-digital technology

As discussed with reference to Table 1, we need to move all our data into files. LTFS allows datatape to move from being non-file-based to being truly a file-based storage medium, but the real problem is all the digital videotape (and digital audio tape = DAT) content.

In the Annual Audiovisual status report published last year, we outlined the three different kinds of digital-to-digital transfer, as follows:

> In an attempt to uphold archive principles and 'save the bits', three cases occur:
> 1) the bits are not available (to the external world); minidisc, Digibeta;
> 2) the bits are available, and a clone is made: CD, DVD, DAT, DV;
> 3) the bits are available, but a clone is not made: this is the case for the D3 preservation project of the BBC

As predicted last year, the issue of digital-to-digital transfer has been very important in 2010. Software to examine read errors from digital sources (DV, D3) has been

[39] http://en.wikipedia.org/wiki/Linear_Tape_File_System

produced (by AVPS for DV tape, and by the BBC INGEX project for D3 and D1 tapes). The INGEX work was started some years ago, but in 2010 the use was extended by Tate Galleries to the D1 format.

Another significant development is a contribution to understanding the problem. AVPS has made a monograph[40] publicly available explaining the digital-to-digital problem using bar-codes as an analogy. A bar code is digital, but it's also on a physical medium that can suffer from various kinds of problems which cause errors in reading the barcode.

This paper has a wealth of pictures showing the results of different kinds, and amounts, of errors for data in the DV format. DV players use error detection and correction to produce valid data, just as for any type of digital storage medium. But then DV players do something extra: when the data cannot be corrected, the player produces a signal anyway, and uses other data in the signal to conceal the effects of the error. The illustrations in the paper give readers a good idea of what kinds of concealment can take place, and their relative success.

The paper then went on to systematically explore damage, playback, concealment and recovery of video and audio from DV-format videotape, including differences between playback decks and differences introduced by various different kinds of playback software.

One might think that recovery of video from a digital format would be easier that digitisation of analogue video, but the paper shows that dealing with errors in digital playback is a real labyrinth, as the BBC has also found in its D3 project.

A conclusion of the paper – and the BBC has separately reached the same conclusion – is that an effective strategy for optimum recovery of video from a digital format would involve stopping, backing up, and repeating the readback process when the hardware sets a flag showing an irrecoverable error – and then patching together the best set of recovered blocks of data.

[40] Barcode Scanners, MiniDV Decks, and the Migration of Digital Information from Analog Surfaces, Dave Rice and Stefan Elnabli, AudioVisual Preservation Solutions
http://www.avpreserve.com/avpsresources/papers-and-presentations/

5 Century Store – shining light into the "digital black hole"

What does it cost to store audiovisual content? The answer comes down to particular cases, rather than generalities – and so PrestoPRIME has produced a tool[41] to provide cost estimates based on user-supplied data defining the 'particular case'.

But there are generalities that have been given wide attention in digital preservation discussions, in particular:

- digital content in general is unaffordable[42]; this is one version of the 'digital black hole' doctrine[43] – the version that says digital is too expensive to maintain and so we're better off with books on shelves, and should stop digitisation projects;
- *Digital film* is far more expensive that analogue film – 11 times more expensive to be precise. This is the AMPAS Digital Dilemma claim[44].

Century Store is a PrestoPRIME total-cost-of-ownership analysis of costs for keeping content for a century. The results show that digital storage for a century, including updating hardware and software and "continuous migration" of files, is not unaffordable, does not cost 11 times as much as 'writing back to film' – and indeed when access costs are considered, digital storage is very likely to be the cost-effective choice.

Despite the digital black hole doctrine, digitisation projects continue and, if anything, increase. The count of Google digitised books is now above three million[45], with another million from the Million Books Project[46]. And in digital film, despite their publication of Digital Dilemma, AMPAS itself is actively pursuing research in digital preservation: the Digital Motion Picture Archive Framework Project[47].

This situation presents a problem: digitisation unaffordable, digital film preservation 11 times too expensive – and yet digitisation continues, and AMPAS studies digital film preservation. Why?

The conclusion we present, with numbers to support the argument, is that the original statements were simply wrong, and people continue to digitise and to use digital storage because they know these statements are wrong.

---

[41] https://prestoprimews.ina.fr/public/deliverables/PP_WP2_D2.1.2_PreservationModellingTools_R0_v1.00.pdf

[42] Jonas Palm: The Digital Black Hole http://www.digiwiki.fi/fi/images/4/4d/Palm_Black_Hole.pdf

[43] Other versions are: 1) No content: for paper we had 'early drafts' and now there is nothing: Without a trace, Malcom Read, Jisc INFORM http://www.jisc.ac.uk/publications/jiscinform/2010/inform29.aspx#withouttrace ; 2) Unusable content: digital content becomes obsolete and so can't be read / rendered / interpreted: http://www.zdnet.com/news/digital-black-hole-threatens-your-documents/150629

[44] The Academy of Motion Picture Arts and Sciences The Digital Dilemma http://www.oscars.org/science-technology/council/projects/digitaldilemma/

[45] http://www.digitaltrends.com/mobile/google-launches-ebookstore-with-3-million-books/

[46] http://www.archive.org/post/305502/over-1-million-digital-books-now-available-free-to-the-print-disabled

[47] http://www.oscars.org/science-technology/council/projects/dmpafp.html

This conclusion is important, because it is uncomfortable to have widely-quoted conclusions that differ from both common practice and common sense.

.5.1 Digital storage will eventually be unaffordable

The argument in the paper The Digital Black Hole (Palm, Footnote 42) is based on looking at the total costs of storage, as follows (p8):

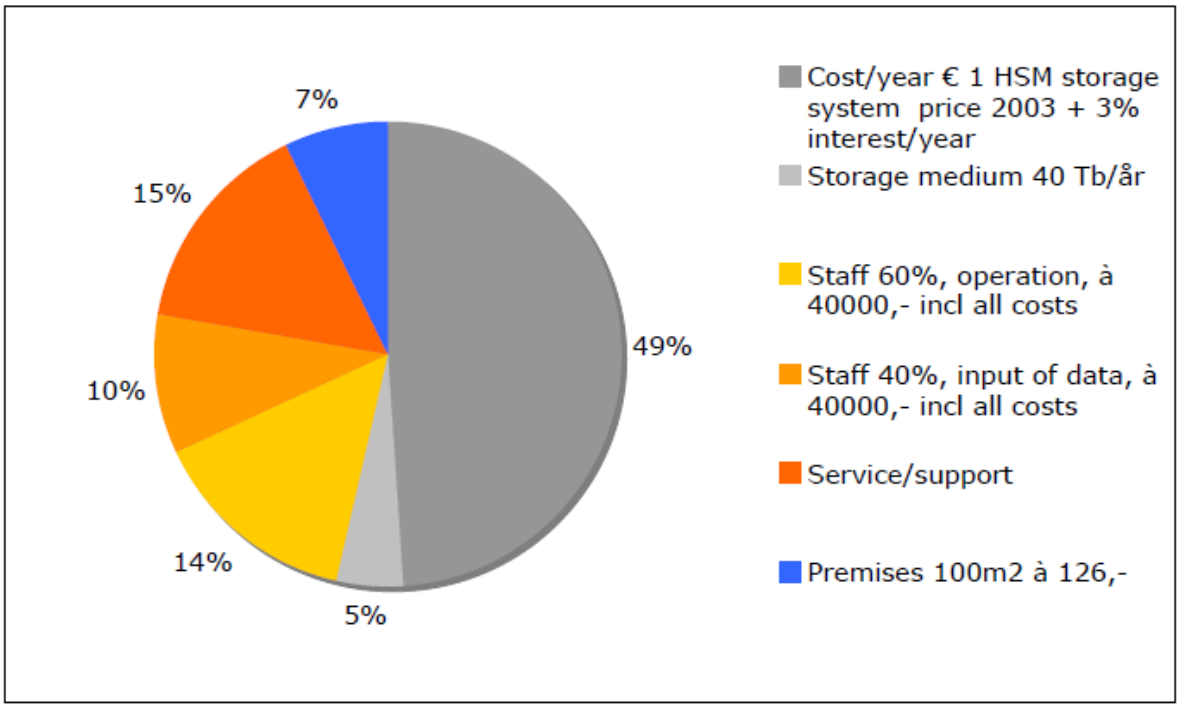


Fig. 10. Division of costs of RA's HSM storage system.

Actual costs are given for all these (Fig 7 in the paper) and they add up to 9.18 Euros per gigabyte (GB), as an average for the five years 2003-2007 inclusive.

The Palm prediction for future costs is based on *negative economies of scale,* the concept that more storage creates even more complexity, and so management costs per unit of storage (GB, TB or whatever) increase over time. There is NO justification whatsoever in the Palm paper for the assumption that storage management will be subject to *negative economies of scale.*

But given that assumption, the consequences are dramatic, as shown in Palm's Figure 12:

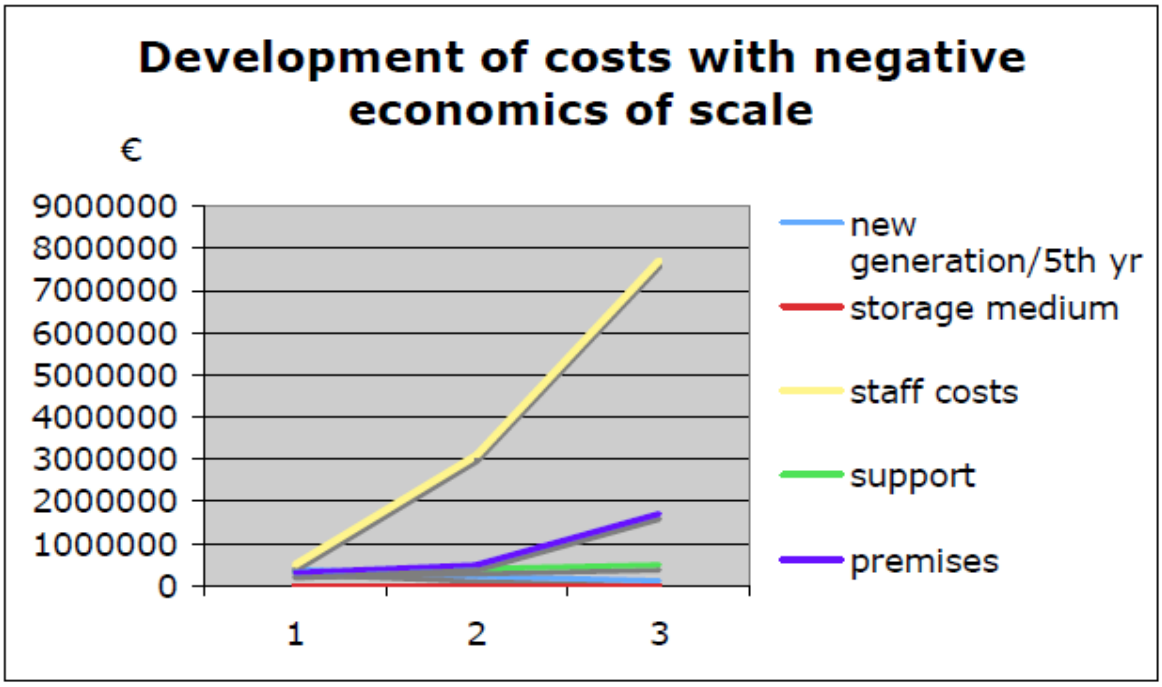


Fig. 12. Probable effects of negative economics of scale at RA in the long run.

Figure 12 has a time-scale in decades. His conclusion is that total running cost of storage will be a huge burden on institutions, and will only get worse. At the risk of being repetitious: **there is no evidence in the paper justifying this expected growth in costs**.

What is the actual evidence? Matthew Addis of the PrestoPRIME project has reviewed all the data that we can lay our hands on, and what we observe is *positive* economies of scale, not negative. This situation makes sense: the terabyte hard drive attached to my laptop needs no more management than the 5 MB hard drive that I first used in the 1980's. In a data centre context, several racks full of terabyte drives occupy the same space, power and staff effort as the same racks full of much smaller drives ten years ago.

For numerical data, there is now experience for up to a decade from data centres, particularly the San Diego Supercomputer Center – which has published detailed storage costs[48], including cost trends over time – from Google[49] and from the Amazon S3[50] cloud storage service pricing.

The conclusion, as already reported by PrestoPRIME in D2.1.1 Preservation Strategies,[51] page 73 is:

---

[48] Moore, R. L.; D'Aoust, J.; McDonald, R. H.; and Minor, D. (2007). Disk and Tape Storage Cost Models. In Archiving 2007. http://chronopolis.sdsc.edu/assets/docs/dt_cost.pdf

[49] Barroso, L. A. and Holze, U. (2009). The Datacenter as a Computer: An introduction to the design of warehouse-scale machines. Google Inc. Synthesis Lectures on Computer Architecture no. 6. published by
Morgan and Claypool.

[50] http://aws.amazon.com/s3/

[51] D2.1.1 Audiovisual preservation strategies, data models and value-chains http://www.prestoprime.eu/project/public.en.html

> "The rate at which the TCO of storage falls doesn't appear to be quite as high as the rate at which raw media costs are falling. SDSC report that the cost of tape halves every three years with the rate for disk being somewhat faster. This is reflected in the rates they charge to customers ($1500 per TB per year on disk in early 2007, $1000 in 2008 and currently $650 as of the end of 2009). Similar trends are seen in the Amazon S3 rate, which in 2007 when SDSC produced their analysis stood at approx $1800 per TB per year (see SDSC paper) and now stands at $1260 per TB per year (if you store over 500TB) and considerably less if you have multiple petabytes to store with the extreme being only $660 per TB for over 5PB of storage.
>
> **It would appear that the real-world TCO for storage halves every 2 to 3 years**, with disk storage falling faster than tape storage. Tape storage using the most recent SDSC figures is still half the cost of disk storage (and that includes 2 copies on tape against one on disk)."

We hope that people will be convinced by our actual reporting of the experience of large data centres.

The difference between negative and positive economies of scale is enormous. With negative efficiency, costs will simply grow and grow, and so anything we commit to today will sooner or later become unaffordable, and overwhelm us.

With positive economies of scale we can do some interesting math: anything that costs less in successive units of time can be added up from now to infinity, and give a finite sum. Non-mathematicians might be puzzled, but simply try adding 1 and ½ and ¼ and 1/8 and 1/16 and so on, and you can readily see that the infinite sum is 2.

Anything that drops by ½ every year adds up to a total doubling of today's unit cost. Further, if the annual cost drops by **1/n** then the sum – from now to forever – is **n**. PrestoPRIME estimates TCO for storage dropping by ¼ per year, hence our rule of thumb that the total cost of storage forever is simply four times the cost for next year!

Even mathematicians express scepticism that storage cost can be either that easy to estimate, or that cheap. But Princeton University[52] is now offering 'forever' storage at a similar multiplier of today's cost: (multiplier of 6). The point is not whether "the answer" is 4 or 6; the point is that there is growing acceptance of the concept of **finite storage cost,** forever, and that the cost is not only finite but really very small – less than ten times the current annual cost.

.5.2 Digital storage is eleven times as expensive as film

It is daunting to disagree with the conclusions of the Academy of Motion Picture Arts and Sciences (AMPAS), and it is probably folly to estimate costs over a century. But as with the Digital Black Hole paper, when there are widely-quoted statements that disagree with published PrestoPRIME findings, the issue can't be ignored.

[52] http://dataspace.princeton.edu/jspui/bitstream/88435/dsp01w6634361k/1/DataSpaceFundingModel_20100827.pdf

AMPAS published the result of a very thorough study into cinema film preservation, called The Digital Dilemma[53]. A wide range of sources was consulted, and it is the only study that has examined in detail the specific and highly complex preservation requirements of the cinema industry. For all those reasons this report has to be seen as an excellent and useful contribution to audiovisual preservation.

But – the one statement in the document that gets widely-quoted is the comparison of analogue to digital storage [Executive Summary, pp1-2]:

> A- "Economic models comparing long-term storage costs of film versus digital materials show that the annual cost of preserving film archival master material is $1,059 per title[54] and the annual cost of preserving a 4K digital master is $12,514,[55] an 11-fold difference."

This quantitative difference then gets magnified by pseudo-mathematical statements on the AMPAS website:

> B- "The long term operating cost of a digital archive, built using traditional IT approaches, is exponentially greater than that of a film archive."[56]

In this section we will examine the calculation in A, and refute it. Nothing can be done about B, except point out that the word 'exponentially' is almost always misused, that "exponentially" is in consequence virtually meaningless – and finally that "exponentially greater" doesn't even have a definition, so it is always meaningless. I guess they mean 'a lot greater' – exponentially being used for emphasis and to sound mathematical. This is an abuse of technical discourse and deserves to be criticised.

But the real problem is A. The situation is that AMPAS have an 'analogue preservation' model for *digital film* which first requires an $80k investment in making an analogue master (actually three physical films: black and white *separations* made by passing a full-colour image through red, green and blue filters). That's a large investment, and the separations could easily last a century, so the AMPAS calculation begins by dividing $80k by 100, to spread the cost over a century.

It is this use of a century in the calculation that forces us to look at costs over a century, despite all the difficulties (and follies) of making such long-term calculations.

Then AMPAS compares the $80/100 (plus $259 for shelf-based storage costs) against the current cost for digital storage of 25 TB of digital content (the amount needed for high-quality digital film) – and gets quote A: $1 059 vs. $12 514, a factor of 11.8.

---

[53] http://www.oscars.org/science-technology/council/projects/digitaldilemma/

[54] Based on a monthly cost of 40 cents per 1,000 foot film reel in preservation conditions plus the amortized cost of film archive element manufacture.

[55] Based on an annual cost of $500 per terabyte of fully managed storage of 3 copies of an 8.3 terabyte 4K digital master.

[56] http://www.oscars.org/science-technology/council/projects/dmpafp.html

The REAL year one costs actually are:

- Analogue year one: $80 000 plus $259 = $80,259
- Digital year one: $12 500

This would leave analogue costing nearly seven times as much as digital, but ONLY for year one. Year two would be:

- Analogue year two: $259
- Digital year two: $12 500 (if this assumption holds true)

So on for a century. Adding it all up gives:

- Century cost Analogue: $80 000 plus 100 x $259 = $105 900
- Century cost Digital: 100 x $12 500 = $1 250 000

In consequence, 1 250 000 / 105 900 = 11.8, meaning we have 'shown' that digital is over 11 times more expensive than analogue. But we've reached this figure by doing two very questionable things:

1. making the most unacceptable assumption about digital storage that we could possibly make, namely that its cost stays constant
2. ignoring access: the cost of actually using what we've stored

The most dramatic fact about digital technology has been Moore's Law[57]: the complexity of semiconductor devices doubles about every 18 months, and has now done so for over 40 years. This author can remember when 'core memory' was roughly $1 *per bit* – and we can now (December 2010) buy terabyte hard drives for $100. By the time this paper gets reviewed and publicly released, the figure will drop to $80, and a year from now it will be $60 – or less.

As already discussed in the previous section, PrestoPRIME estimates the total cost of storage – forever! – as four times the current annual cost.

The PrestoPRIME calculation for comparing analogue and digital would be:

- Century cost Analogue: $80 000 plus 100 x $259 = $105 900
- Century cost Digital: 4 x $12 500 = $ 50 000

What if Moore's law doesn't go on for 100 more years? The practical answer is: it hardly matters, providing it goes on for another 20 years. In 20 years, dropping by 20% every year, the cost of storage will be about 1/80th of the present cost. If it then stays at that level for 80 years, the calculation becomes:

[57] http://en.wikipedia.org/wiki/Moore's_law

Century cost Digital: dropping by 20% per year for 20 years = $49 217
plus another 80 years at $180/yr = $14 400
Century cost: $63 617

The above calculations deal with problem 1: ignoring Moore's law. An equally serious issue is problem 2: access. Analogue separations in a box on a shelf have no access. They need to be scanned to be re-used in digital production, or even to be viewed in digital cinemas. The current cost for high-resolution scanning is NOT given in the AMPAS Digital Dilemma report. Scanning costs have been coming down for the last decade, but one driver has been the desire of archives to move away from film, to move where the market is. Once that process has largely been accomplished, film scanning equipment may become very exotic and therefore costs could rise.

Sun estimated the current cost of high-quality rescanning as $60k[58]. Whether that drops significantly or not, the conclusion is that access to analogue separations is expensive, and will remain so. A proper total cost of ownership model for preservation of digital cinema should include a model for access (uses per century), and costs added in to cover that projected use.

For instance, if a film is to be used twice in a century, at a cost of even $20k per scan, the century cost comparison becomes:

Century cost Analogue: $80 000 plus 100 x $259 = $105 900
Access: 40 000
$145 900

Century cost Digital: somewhere between $50 000 and $65 000

Note that nothing was added to the digital cost estimate for access, because nothing needs to be added. The digital storage is on files, and the cost model includes keeping the storage current (migration) and keeping the bits correct (curation). Content on digital storage is access-ready, by definition.

The conclusion is: analogue storage of digital cinema is not a factor of 11 cheaper; it probably is more expensive, and when cost of access included it is very likely to be much more expensive.

.5.3 Why we don't minimise total-cost-of-ownership

Total Cost of Ownership (TCO) is not everything. We are all under budget pressures, and there is the danger of making decisions based just on cost. Archivists should remind ourselves – and anyone who will listen to us – that there are other dimensions that matter.

Here are three decisions that we would have to make if TCO were the only criterion:

[58] Archiving Movies in a Digital World, Dave Cavena et al, January 10, 2007; http://sun.com/storagetek/disk_systems/enterprise/5800/ArchivingMoviesinaDigitalWorld.pdf

- **Only make one copy**
  “One copy is no copy” is the archivist's rule of thumb, but we could immediately save 50% by only having one copy. Yet we do try to have two copies, whenever we can, because of all the risks associated with only one master.
    - Conclusion: TCO isn't everything: **risk** must also be considered

- **Do nothing (watch it rot)**
  We could make massive TCO savings by not doing any preservation or even routine maintenance, or even paying for sensible storage conditions. But then the archive disappears, and with it whatever benefit the archive had ever brought, to anybody.
    - Conclusion: TCO isn't everything: **benefit** must also be considered

- **Make a DVD (from your deteriorating 2”, DigiBeta or film)**
  Clearly the cheapest way to digitise videotape or film is to make DVDs. They have a quality as good a current television transmission standards. This approach would violate all the professional practices ever established for either preservation or for professional access and production, but it would save a lot of money. In the library world, it would amount to replacing parchment originals with (very ordinary) photocopies. But not every document is a parchment illuminated manuscript of obvious value, and making a mediocre-but-cheap copy could be seen as better than doing nothing.

  It would be better than doing nothing, and cheaper than 'doing it right' – but it would also be an acceptance, forever, of dramatic loss of quality; so we would resist it.
    - Conclusion: TCO isn't everything: **quality** must also be considered

Cost, risk, benefit and quality: four dimensions to balance, and there are no doubt other considerations. These are sufficient to remind ourselves that TCO isn't everything.

The conclusion is: for all archive decisions, from investment in storage media to use of data reduction (compression), there is unlikely to be a single dimension such as cost on which to base decision making. Cost, risk, benefit and quality (at least) must be balanced.

6 Relevance of Digital Preservation to Broadcast Archives

Audiovisual content is produced by many institutions (research, universities, the arts, medicine, defence, satellite imaging, undersea and geological exploration ...) and held in many places (museums, galleries, libraries, archives, local history collections, oral history collections, government records offices, universities, national heritage institutions) in addition to broadcasting. However broadcasting does produce – and hold – about 40% of all audiovisual content[59].

In this section of the Status Report, we go through the existing digital preservation technology and the work of PrestoPRIME, and look at its relevance to, and potential impact for, broadcasting.

The *Relevance to Broadcast Archives* section covers:

- Where Digital Preservation Technology is Needed
- Fundamental Differences of Digital Content
- A Strategy for Digital Preservation
- Standards: OAIS, Europeana, Metadata
- Processes: Digitisation, Digital preservation, Management
- Tools
- Public Value
- Steps Toward Creating a Broadcast Digital Archive

.6.1 Where Digital Preservation Technology is Needed

All broadcasting and indeed all audiovisual production is moving from physical media as carriers of content, to file-based content. There is long-term value, to the broadcasters and to the public, in preservation and above all in re-use of broadcast output, and so the major broadcasters have maintained archives (on shelves) for many decades. How can file-based content be preserved and re-used equally successfully – or better? PrestoPRIME is the effort of five major European broadcast archives (with a range of partners) to developing technology in the general area of *digital preservation* – keeping those files, or at least the content within those files, preserved and usable for so long as the archive requires.

This section of the Annual Status Report summarises the major digital preservation problems faced by broadcasting, concentrating on the solutions and benefits arising from PrestoPRIME. The problems can be divided into five major areas:

Strategy – the overall approach to preservation and re-use of audiovisual content: a preferred path through the options of encoding, compression, wrapper formats, metadata types and management, storage, outsourced services and obsolescence (of everything).

Standards – a preservation system for audio and video files, based on compliance with the OAIS model (and standard); the system includes detailed specification of the *information packages* needed to preserve digital content in a structured and assured fashion, following the OAIS principles. The packages themselves follow digital

[59] TAPE final report: Lusenet: Tracking the. reel world : A survey of audiovisual. collections in Europe, http://www.digiwiki.fi/fi/images/c/c4/Tracking_the_reel_world.pdf

preservation standards. In consequence, PrestoPRIME is defining the de facto digital preservation standard for audiovisual content.

Processes – workflow for effective digitisation and digital preservation. This area addresses the European challenge of at least doubling the current digitisation rate, through developing tools for *computer*-supported quality control. The need for a much greater rate of audiovisual digitisation has been highlighted as long ago as 2002[60], when the Presto project showed that new material arrived at audiovisual archives at four times the rate at which current content was being digitised. PrestoSpace extended these estimates, predicting the loss of between 40% and 70% of current audiovisual content *in formal, curated collections,* unless there were an fundamental change of approach[61].

Tools – while PrestoPRIME is developing an overall preservation system (so that end-to-end workflow can be integrated and tested), it recognises that what is most needed are ways to fill the gaps and augment existing systems, as few if any institutions will replace existing systems and workflow just to improve their digital preservation functionality. Accordingly, all the PrestoPRIME technology is organised as a component-based architecture following open standards – to maximise the opportunity for these tools to be used in other systems.

Public Value – many audiovisual collections have major plans for public access, but there is no experience, anywhere, of how to make a million hours of content accessible to the general public. Europeana, the main European cultural heritage portal, provides a proving ground for access technology, including Thought Lab, a variety of semantic access tools[62] developed by the Free University of Amsterdam. The PrestoPRIME partners are also launching PrestoCentre [Section 1.4, this report], an advisory service to 'outlive PrestoPRIME' and support saving the European audiovisual heritage.

.6.2 Fundamental Differences of Digital Content

Broadcast archives have been building archives of re-usable content since around 1930[63], and have shown the rest of broadcasting the value of holding and re-using programmes and related materials. These archives have world-class knowledge about how to run shelf-based collections, but very little experience of running file-based archives – and the experience they do have is based mainly on archiving of text files (electronic documents).

There are many differences between audiovisual files and text files:

1 **Complexity:** not just the range of file formats, but the fact that a file can contain (wrap) many elements: video, multiple sound tracks, subtitling, time code, metadata. Added to this are the many types and qualities of encoding (MPEG 2,

[60] Presto project deliverable D2 Archive Preservation Survey http://presto.joanneum.ac.at/projects.asp#d2
[61] PrestoSpace Final report on users requirements (D2.1) http://prestospace.org/project/deliverables/D2-1_User_Requirements_Final_Report.pdf
[62] http://www.europeana.eu/portal/thoughtlab.html
[63] BBC Gramophone Record Library Valentine Britten, BBC; Published 1963 http://openlibrary.org/books/OL5575542M/BBC_Gramophone_Record_Library

MPEG 4, JPEG, JPEG2000, MJPEG, AVI, MOV, WAV, MP3 and dozens more[64]), the fact that multiple versions (production quality, transmission quality, browse quality) have to be managed as a unit – and finally the issues of Digital Rights Management (DRM).

2 Audiovisual content represents an analogue **signal**, and so has dimensions of frequency response and signal-to-noise ratio that determine how faithfully a sound or image is captured or reproduced. In addition to "preserving the bits" there is always the added dimension of whether "the bits" have preserved the signal.

3 **Compression:** audiovisual signals contain redundancy, and so for decades[65] people have pumped high-bandwidth signals through low-bandwidth channels by manipulation to remove the least significant (hopefully) information. Only lossless compression is used on text, but audiovisual content gets subjected to processes that throw away parts of the signal: lossy compression. The effect, years later, of repeated application of various kinds of lossy compression is a hazard unique to audiovisual content.

4 **Size:** "above four gigabytes, everything breaks[66]" is not absolutely true, but it is true that storage, systems, networks and applications are stressed by large files. Four GB was the memory limit of 32-bit processors. Personal computers are now mainly 64-bit, but many embedded processors that control everything from coders to routers to storage devices are not 64-bit – and many applications have limitations of their own (analogous to the Year 2000 problem) which are independent of hardware and which can kick in at less than four GB. An hour of standard definition video at full quality (uncompressed) is about 100 GB.

5 **Time:** audio and video have a time dimension. Metadata and applications need to understand that dimension, so documentation and access can "point to the right place" rather than simply dealing with an audiovisual file as a unit (lump).

6 **Resilience:** for decades broadcasting has coped with errors. A glitch from an analogue VTR (videotape recorder) could be smoothed out using a time-base corrector; a glitch in playback using a digital VTR could be concealed through repetition of an adjacent line, or even an entire frame. Systems for files use built-in error detection and correction, and if that fails, the file as a whole can be rejected, generating an error message saying something like "file cannot be opened". Because video is highly structured (into lines and frames), there is potential for playback despite errors – if only the IT systems would pass 'the good bits', plus an indication of where (along the time dimension) the error occurred.

7 **Access:** four requirements for time-base access are fundamental to all capture, storage, archiving, access and re-use of audiovisual content

---

[64] FFMPEG supports 22 families of coded, each with several varieties; http://en.wikipedia.org/wiki/FFmpeg#Codecs

[65] Since the *vocoder* (Homer Dudley, 1936) for speech and the image compression used for facsimile equipment (1950s) – based on the Baird system of image scanning (1920s).

[66] Matthew Addis, IT Innovation, University of Southampton; project Avatar-m: http://www.avatar-m.org.uk/

7.1 *Granularity:* division into meaningful parts, so the item can be represented in some visual way (e.g. by keyframes or a storyboard), supporting navigation (getting to the right place)

7.2 *Click and play:* playback from any point, so the user can click on a keyframe (or an audio equivalent) and immediately get to the desired place in the content

7.3 *Citation:* creation of a pointer (to a specific time point within a file associated with a permanent URI) that can be put in an email, website – or Ph.D. thesis.

7.4 *Annotation:* adding commentary at a citation point. If the commentary can be accessed by other people, there is then a basis for building a community or social network around the annotations and the annotation process.

Audiovisual archives now have to develop secure ways to manage and keep audiovisual files for the long term, meaning at least for decades. This new *digital archive* has to be developed, and thrive, in a business where "a few days" is considered a long time.

The major experience in digital archives lies with libraries: national, research, academic. They have been building digital collections, and developing *digital library technology*, since the mid 1990's[67]. More recently, libraries, archives and government bodies concerned with the permanence of digital materials have moved on to develop *digital preservation technology*, with the Open Archive Information System[68] (OAIS) model (made an ISO standard in 2003) and the development of the concept of *trusted digital repositories*[69] – meaning trusted to keep things long-term. Unfortunately for audiovisual content, the work on digital libraries, archive and repositories and the related work on digital preservation has concentrated on the main business of archives and libraries: *documents containing (mainly) text and still images*. Sound and moving images have all the additional issues listed above, with requirements that have not been addressed by mainstream digital preservation approaches. The result is that many tools developed for file-based archiving and preservation only support document and image formats, and certainly don't support MXF[70].

A related problem arises within broadcast archives: their 'world' is the world of broadcasting, not national and university libraries and archives. Their technical and IT staff and systems are largely unaware of the technology, standards and systems developed for digital archiving and preservation. There is a 'two worlds' problem: those who know and use MXF, and those who know and use OAIS. Broadcast archives are caught in-between.

---

[67] The US Digital Library Initiative dates from 1994: http://dli.grainger.uiuc.edu/national.htm
[68] OAIS: http://www.ukoln.ac.uk/repositories/digirep/index/OAIS
[69] Trusted Digital Repositories: Attributes and Responsibilities; an RLG-OCLC Report. RLG Mountain View, CA May 2002 http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf
[70] None of the standard digital library tools listed here supports MXF: http://code.google.com/p/fits/

PrestoPRIME[71] was created by broadcast archives that *do* have knowledge of the digital library and digital preservation world, in order to bring the disjoint technologies together – at least in the area of digital preservation. The work and results of PrestoPRIME are presented in the following sections.

.6.3 A Strategy for Digital Preservation

The PrestoPRIME principles for preservation of standard-definition video are very easy to state:

- Save the originals
- Digitise at full quality (according to the SDI[72] standard)
- Save in a professional wrapper (e.g. MXF)
- Save using open standards (e.g. Linux TAR files on datatape)

These principles should be familiar to the broadcast archives, because they come from the BBC R&D Ingex[73] project. Ingex sets a goal: *preserve the uncompressed, full-quality signal*. The problem PrestoPRIME addresses, and which broadcast archives face in many areas, is **how** to get to this goal. PrestoPRIME also adds documentation and compelling examples supporting **why** it is important to get to the goal. PrestoPRIME adds the how and why, and has the task of communicating that not just to broadcast archives, but to the whole audiovisual community.

.6.3.1 How to get to the goal

Uncompressed digital video is not the usual starting point. Archives start with analogue (e.g. 2", 1", U-Matic, BetaSP) or compressed digital (e.g. DigiBeta, the various kinds of DV) formats. Current production may be based on MPEG 2(D10) or MPEG 4 (AVC) compressed files. Digital cinema has standardised on distribution of MXF files wrapping a (lossy) JPEG2000 encoding – meaning film archives are faced with receiving 'originals' that are already compressed. Many "capture formats" – including virtually all HD capture and production formats – are compressed.

As long as the compressed format, in its wrapper, is usable – there is no problem. How long is that? PrestoPRIME has produced a flowchart for decisions about keeping vs. moving on, and in the detailed explanation we give references to the major sources of information about wrapper and encoding obsolescence – and other risks and dependencies. These decisions are vital to broadcast archives: we know how to save uncompressed SD, but the situation for HD, for digitised film and for 3D is not clear, and will certainly involve making decisions about compressed files – following PrestoPRIME guidance.

The following diagram is the generic flowchart. It starts from a position where the master material is compressed video, and has the following key stages:

---

[71] PrestoPRIME is an EC-supported research project; the principal partners are BBC (UK), B&G (Netherlands), INA (France), ORF (Austria) and RAI (Italy); each is (or holds) the major public service broadcast archive in its respective country. http://www.prestoprime.org/ http://wiki.prestospace.org/pmwiki.php?n=Main.PrestoPRIME

[72] http://en.wikipedia.org/wiki/Serial_digital_interface

[73] BBC Ingex Archive http://ingex.sourceforge.net/archive/

- Assess: check the encoding and wrapper (format) for risks. In the case of broadcast archives, this check is whether the production and exploitation tools in current use, still work on this format.
- If so: “archive temporarily”. **Temporary archiving** is a phrase coined by PrestoPRIME, to highlight the fact that the goal has not been reached (hence temporary) -- but the format being held in the archive is usable and on the path to that goal (hence archiving).
- If not: something has to be done. Generally digital preservation technology is based on migration from obsolete file formats to new ones, but PrestoPRIME includes the **multivalent** approach[74] developed by University of Liverpool: interpreters for particular file types, written in Java, and so in principle executable indefinitely on a Java Virtual Machine.
    - The multivalent path allows content in an obsolete format to still be used, for so long as the dedicated multivalent player is usable. The output of multivalent could be a rendering through a browser, or it could be another file – in a standard or canonical form – produced using a Java-based tool. The idea is that the Java-based tool can be kept viable when other tools have become obsolete.
    - The ad-hoc path could be finding a different tool (e.g. editor) that still works on the old format.
    - The three migration options are colour coded. Green gets to the goal, yellow is on the path to the goal, and red drops off the path because going from one lossy compression to another is NOT maintaining quality: it is sacrificing quality because such a migration is equivalent to a generation loss in the analogue world.

PrestoPRIME has detailed case studies for the major archive and production formats, and *worked examples* of the flowchart, showing actual recommended practices for specific formats and situations[75].

[74] http://multivalent.sourceforge.net/
[75] D2.1.1 Audiovisual preservation strategies, data models and value-chains http://www.prestoprime.org/project/public.en.html

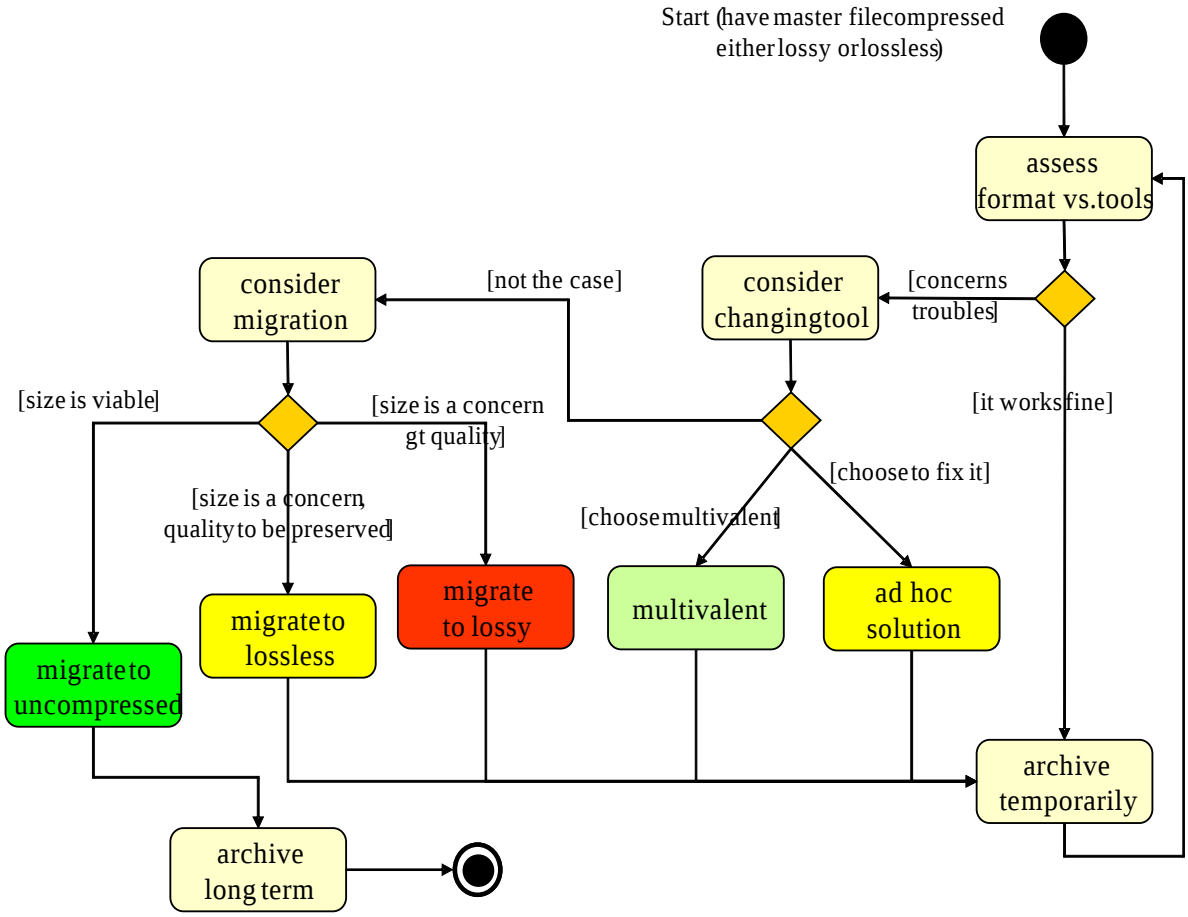


Figure 1 - Decision flow chart for managing files (see explanation above for details)

.6.3.2 The need for uncompressed content

A “compressed original” is not an immediate problem in itself. If that was the production format, then it represents broadcast production standards and should be a good representation of the visual signal – and probably contains an uncompressed audio signal. The problems are all about what comes next, over time:

- Obsolescence:
  - Of the wrapper or the encoding, both of which may be superseded; examples are the progression from MPEG 1 to 2 to 4, or from AVI to MOV or MXF.
  - Of the browse-quality version of the same content.
- Resilience: resistance to errors or corruption. The IT industry emphasises its low error rates, but small numbers (low error incidence) become large ones as broadcast archives moves wholesale to file-based production, meaning huge files and large numbers of huge files.

Why uncompressed?

- There is no codec to ever worry about. Codecs add complexity, they can become obsolete, they may have licensing issues, and they add time and computing requirements to any use of the material. Uncompressed material still needs software that can understand it (edit it, make a viewing proxy from

it) – but it will never need that software to also understand a particular coding method, and use time and resources to decode.

- The video content is guaranteed to be of production quality. All risks associated with whether or not DV or MPEG 2 or MPEG 4 at one of a dozen possible profiles – is good enough to for production-quality reuse are eliminated at a stroke, forever. The *signal* in the file could still be of poor visual or aural quality, but at least everyone can be sure that 'the bits' are as good as they could ever be.
- With no encoding, there is no possibility of multiple encodings. Stringing one decode-encode after another (where each on its own has no perceptible effect, and particularly when the coders use different methods) can result in grave or even catastrophic degradation. One very common characteristic of digital processing in general is that problems tend to get concealed until they become overwhelming, resulting in moving suddenly from OK to unusable.
- There is no generation loss. Moving from one lossy codec to another over decades implies a continuous drift away from 'the best', with no way back. Once an uncompressed version has been saved, that will never drift, and will be as close to permanent as can ever be achieved.
- The processing to make new browse or web copies is minimised – simply because the decode stage (required by compressed material) is eliminated, forever.
- The uncompressed version is canonical – it is 'the standard form'. If the bits are preserved, then whatever becomes of future technology, a signal can be recovered from any uncompressed audio stream (trivially) and from any uncompressed video stream (not quite trivially, but only a small number of options are possible).
- The uncompressed version is simple. Codecs are complicated, and digital preservation of an encoded signal requires archiving not just 'the bits', but the whole environment (codec and player software, plus whatever dependencies are introduced by such software) also needs to be both documented and archived. Uncompressed signals can survive without any such complexity.
- The uncompressed version is robust. The significance of bit-rot (bits flipping while on storage) and read-back errors (not getting back from storage the same bits as were sent in) increases as files get larger, as more content is processed as files rather than as tape, as storage devices get larger (meaning greater loss per device failure) – and the significance of errors increases over time as more and more data is stored and more and more data is moved from device to device and from system to system.

Error rates for storage devices of 10^{-13} become very significant when the amount stored and moved becomes 10^{+15}. Device failures and system corruption errors have occurred, and their incidence can only grow. Storage devices themselves cannot be expected to reduce their errors below current levels, based on performance over the last several decades. Storage systems are increasing in complexity, moving from simple file-management to media asset management – often combined with other complexity such as encryption and digital rights management. PrestoPRIME has

researched the risks[76], but the bottom line is clear: errors happen, and compressed files magnify errors while uncompressed files localise errors.

Here is an example of many errors in an uncompressed file, vs. a single bad byte in a compressed file:

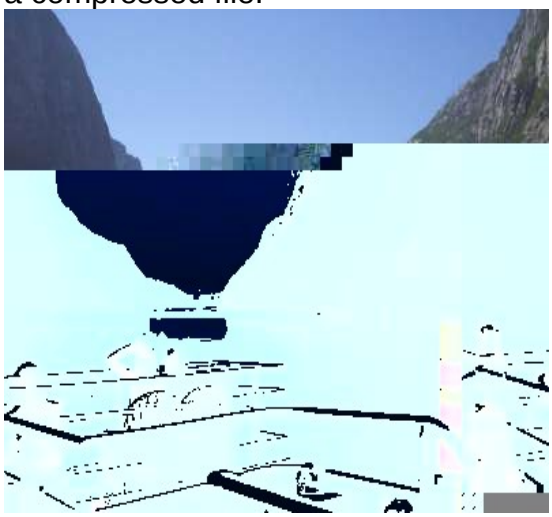


Figure 2 Compressed JPG file, 14k, 1 error

Figure 3- Uncompressed BMP file, 350k, 1400 errors

.6.4 Standards

.6.4.1 OAIS

The major digital preservation model – and standard – considered by PrestoPRIME is OAIS, the Open Archive Information System[77]. One reason for paying attention to OAIS is because it is the only standardised approach. There is plenty of experience across the IT industry of *best practice* for keeping data, but such information is a weak basis for creation of a digital archive: the detail is spread out across various bodies, codified in various ways, available in various ways – and generally too vague and diffuse to form the foundation for a digital archive.

OAIS by contrast is a single thing: the OAIS standard. It arose from the concerns of the US Space Programme (NASA) regarding long-term preservation of their material, and has subsequently had international input and support from government agencies (particularly archives and libraries) and universities (again, particularly libraries) – hence becoming an ISO standard.

Why should broadcast archives bother about OAIS?

- The only game in town: OAIS is the only standardised approach to digital preservation – the common ground that libraries, archives, hardware and

[76] Addis, M., Lowe, R., Salvo, N. and Middleton, L. (2009) RELIABLE AUDIOVISUAL ARCHIVING USING UNRELIABLE STORAGE TECHNOLOGY AND SERVICES. In: Conference of the International Broadcasting Convention (IBC 2009), September 2009, Amsterdam, Netherlands. http://eprints.ecs.soton.ac.uk/21066/

[77] http://www.ukoln.ac.uk/repositories/digirep/index/OAIS

software systems developers, vendors and government can agree on and build on.

- Governance: OAIS gets incorporated into government policy[78], which eventually will result in regulatory bodies asking broadcast archives to account for their digital preservation approach. OAIS will be expected to be part of the answer, and may even end up as being a required part of any acceptable answer.

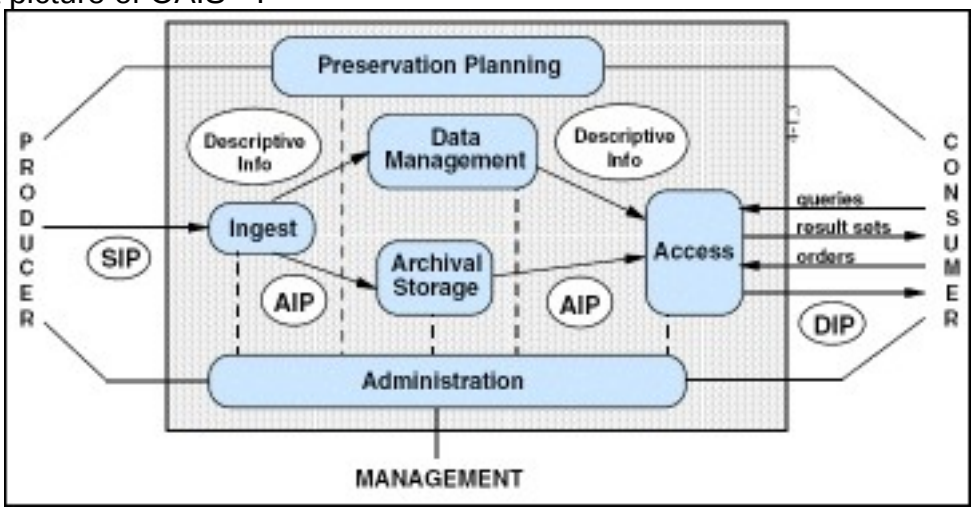
- Archive policy: major libraries and archives have already identified the OAIS approach as necessary, and broadcast archives that aren't aware of OAIS will risk being considered substandard.

The problem with OAIS is that, despite being an ISA standard, it is basically an approach. It does not reduce to numbers and recipes, so nobody can say exactly what OAIS compliance means. The role of PrestoPRIME is twofold:

- Gathering information on best practice in audiovisual archiving.
- Developing our own PrestoPRIME implementation of key components of the OAIS approach, *aimed specifically at professional broadcast files.*

.6.4.2 What is OAIS?

Here's a picture of OAIS[79] :



The picture shows a box which has a number of elements and paths, and mysterious circles labelled SIP, AIP and DIP. Important points about OAIS include:

- There is only one way in: a Submission Information Package (SIP) is the ONLY thing that comes in.
  - The first implication is that this package has to be complete: everything that anyone, anywhere, ever will need to know about the content – is meant to be part of the SIP. A tall order. PrestoPRIME is developing SIPs for standard broadcast files, and will have examples of how they have been created and used in various real broadcast environments (ranging

[78] for instance, OAIS compliance by the UK Data Archive and The National Archive (of the UK) has been evaluated: http://www.nationalarchives.gov.uk/news/stories/79.htm
[79] from http://www.preserv.org.uk/

from WGBH and WNET in the USA, to the PrestoPRIME partners in Europe).
    - The next implication is that a SIP is complex, containing already complicated wrappers such as MXF and adding provenance, technical description, context and usage description, required rendering software and everything else needed to be 'complete'. So either a SIP is a file which can function as a 'wrapper of wrappers' – or a SIP is not a single file at all, but a package as the name implies. A package is an organised and documented *set of files*. A set of files is not a standard unit in IT practice. There are *folders*, but no standard for what a folder is and how it behaves. To make a SIP a precise thing, the METS standard has been used in all known OAIS implementations, and PrestoPRIME has a METS object containing what we deem necessary to make a *broadcast SIP.*
- There are many ways out, but the SIP never comes out. Indeed, the SIP isn't even stored. A SIP is converted by an i*ngest* process to an Archive Information Package (AIP), and to internal (to the OAIS system) metadata for managing the AIP (forever). The AIP is placed in 'archival storage' – whatever that is – and the metadata goes into the Data Management database.
- The AIP never comes out either. So we presume it stays in the OAIS forever – because what does come out are Delivery Information Packages (DIPs). These are proxies: copies of the AIP that are not (necessarily) exact copies. They could be reduced in quality (a Browse DIP) – or could be an extract (and Edited DIP) or any other manipulation.
- There are two functions that operate the OAIS: *administration* and *preservation planning*.
    - An OAIS with just *administration* would be a digital library: something that holds what it's given, and supplies copies or versions as needed. The administration function uses the Data Management data to find things, and supply user needs (for DIPs).
    - The *Preservation Planning* functionality is the most interesting part, because here are the functions needed to keep the content working (viable; renderable) despite obsolescence of the formats of the original data from the SIP.

PrestoPRIME is coding actual SIP, AIP and DIP packages, taking note of best practice from the library and archive world – and of the very limited examples from the broadcast world (basically one project and WNET and New York University[80]).

In addition to the Information Packages, PrestoPRIME is looking at the functional parts of the OAIS model:

- *Ingest*: this strips out the metadata, and sends it to Data Management. Immediately there is a problem, because either the SIP contains completely standardised metadata, or there is a need to map non-standardised metadata into a common schema. PrestoPRIME has analysed gaps in commonly-used audiovisual metadata[81] and is developing a mapping tool to solve the problem of getting metadata from various kinds of audiovisual files into a single, common Data Management system.

[80] Preserving Digital Public Television http://www.thirteen.org/ptvdigitalarchive/
[81] D2.2.2 Metadata Models, Interoperability Gaps, and Extensions to Preservation Metadata Standards http://www.prestoprime.org/project/public.en.html

- *Data Management*: this is the catalogue, the inventory of the OAIS. It requires a data model, to do anything. PrestoPRIME has a data model under development, and also ExLibris is extending their OAIS-compliant data model to accommodate broadcast content.
- *Archival Storage*: this component is largely undefined in OAIS[82].
    - Separation of storage management from storage devices: a storage system needs a mechanism so that the physical storage can be changed (completely) without affecting the logical storage system. Avatar-m has this technology.
    - Permanent identifiers: files have names and locations. There is an inevitable tendency to think that a digital archive is just some sort of database holding lots of names, and lots of paths. The opposite is true:
        - Physical paths should never be part of the identity of an object, because physical paths HAVE to change, over time;
        - File names also should never be part of the identity of an object, because *preservation actions* such as migration from one file format to another would change file names, or at least filename extensions.
- *Administration:* ExLibris have an operational example of a working, complete, OAIS compliant digital library and digital preservation system, and full information on this is available to broadcast archives through PrestoPRIME.
- *Preservation Planning*: PrestoPRIME has a formal strategy for digital preservation, based on the flowchart in Figure 1, above. The components (multivalent, migration) are being developed as tools within a *preservation toolkit*[83].
- *Access*: PrestoPRIME will define and produce at least two kinds of DIP, one for full quality content and one for browse or web quality content. As most of the world's content is not in an OIAS, PrestoPRIME is also has a major component of work on access that is independent of OAIS.

.6.4.3 Europeana

Europeana is the European portal for digital library content. It is not a standard, but it uses and develops standards, and is at least as important to cultural heritage digital library content (in Europe) as is OAIS. Europeana is a partner in PrestoPRIME, and PrestoPRIME is a partner in the cluster of projects around Europeana that are defining standards and processes, developing technology and aggregating content for the Europeana portal.

Digital libraries began with universities, and people in universities do research which they communicate through formal papers. These papers get printed (eventually) in journals, but for roughly 20 years scientific discourse has been moving more and more to electronic documents, shared electronically. How can a researcher find a relevant document, among the hundreds of thousands produced globally every year? The answer is metadata, but the metadata needs to be collected and made

[82] The description is two pages (out of 148 total) in the basic OAIS document, and those two pages refer to functions to be carried out, with nothing about how to carry them out. http://public.ccsds.org/publications/archive/650x0b1.pdf Jan 2002; pp 4-7 and 4-8

[83] D3.1: Design and Specification of the Audiovisual Preservation Toolkit http://www.prestoprime.org/project/public.en.html

searchable. Search engines now provide a non-systematic method of finding information, but beginning 20 years ago a systematic approach was developed based on getting academics to use a common descriptive standard, following the Open Archives Initiative[84] (OAI, which is nothing directly to do with OAIS). OAI metadata on internet accessible documents can be harvested, and used to build catalogues (directories, databases) of content in various subject areas. There is a defined process for collecting the metadata (meaning literally a defined way to query servers connected to the internet and collect the responses). The method is OAI-PMH[85], the Protocol for Metadata Harvesting (of OAI metadata).

Europeana uses OAI-PMH to collect cataloguing information on several million objects, with growth plans for reaching ten million items during 2010[86]. PrestoPRIME has a role here, because, as was the case with OAIS, technical and IT people in broadcasting tend to have no knowledge or experience of OAI-PMH, which in turn means that they have no understanding of how to make their content available through the Europeana portal.

Does Europeana matter to broadcasters? Only to the extent that broadcasters want to be part of public access to digital library content. There is a choice: Europeana is a portal to serious cultural heritage content: digitised books, photographs, records and related text and image material from national libraries and other 'cultural heritage institutions'. Broadcasters can be in, or out. What public service broadcasters cannot do, legitimately, is ignore a cultural heritage initiative as major as Europeana – and then complain about being perceived as 'just entertainment' rather than as part of 'serious cultural heritage'.

.6.4.4 Metadata

Finally, PrestoPRIME is about standards for metadata. If there is one thing the world does not need, it is more standards for metadata, so what is PrestoPRIME doing here? The work of PrestoPRIME will be to actually *use*[87] standards, to show how metadata can be:

- found and understood,
- used to find content and then to navigate around it (with time-based metadata),
- saved in an OAIS,
- used in an OAIS to support preservation and access.

The *use* of metadata standards will be defined in the following areas:

Access: PrestoPRIME has explained (D6.2.2[88]) how Europeana uses the OAI-PMH to collect metadata; there is also explanation of how to create OAI-PMH metadata. PrestoPRIME also supplies information on three difficult areas:

---

[84] http://www.openarchives.org/
[85] http://www.openarchives.org/pmh/
[86] Europeana reached 10,777,149 objects at July 2010: Europeana Highlights - July 2010 http://version1.europeana.eu/c/document_library/get_file?uuid=ea9f2b40-1730-4ab0-a3d6-5b3f2a051edb&groupId=10602
[87] It is the author's view that many metadata standards have had more effort put into their creation than into their use. PrestoPRIME should help redress the balance, by concentrating on use.
[88] PrestoPRIME D6.2.2 European Digital Library implementation guidelines for audiovisual archives

- on how to use semantic metadata from completely different ontologies[89]
- a tool for mapping (to a common schema) the metadata found in typical broadcast file types (also in footnote 89)
- A critical review (footnote 81) of the metadata standards available for audiovisual content, showing their gaps and what can be done about them.

PrestoPRIME will show best practice for how time-related information (e.g. timecode of various sorts, subtitles, time-based annotation) is carried in standard broadcast files – and how this information can be used to access material at a desired time point. PrestoPRIME will explain how time-based information can be extracted from broadcast files, made accessible by OAI-PMH, and end up in digital repositories (or portals e.g. Europeana) in a usable way.

There is a serious issue for broadcast archives, because time-based access can be locked into an application, rather than being generally supported (by time-based metadata being open, standardised and readily available to all applications). If the only way to navigate content is via one brand of media asset management system, or via one kind of file in one brand of edit system, then usage is restricted in the present (to users of those systems) – and compromised in the future (time-based information inaccessible, unusable or lost). Time-based metadata has to be managed within broadcast archives is a way that is accessible to all applications (from high-end edit to public web access), and that can be stored as an open standard in an OAIS. PrestoPRIME will demonstrate how this capability is achieved.

Use: access is about finding content, and getting to the right place. Use is about pointing to the content: citation. While citation is a word associated with formal documents, the ability to cite – to point to something, know it is there, know it will stay there – is fundamental to any concept of a digital repository. PrestoPRIME will show how to point to content, including how to point to a particular place along the time dimension, in a way that aligns with standards and best practices. PrestoPRIME is working with the relevant W3C committee in this area[90].

Provenance: this is an area where PrestoPRIME is creating a standard, but a needed one: creation, storage and use of *fingerprinting* as a method of identification of video content. However, creation of the standard is the minor issue; the real work is developing the fingerprinting technology itself. The project partner INA already has a commercially-available fingerprinting system[91]. For PrestoPRIME, INA has developed a lightweight system of fingerprinting[92], where the fingerprint information can be computed with much less computational effort. PrestoPRIME is developing the PrestoCentre, a European network of national *competence centres*[93], to support

[89] Deliverable D4.0.1 "Initial version of metadata conversion and deployment, vocabulary alignment, annotation and fingerprint computing services". (pdf, v1.00, RE, 03/09/2010). This document describes initial software prototypes of the metadata mapping and validation services, the video tagging game and the audio and video fingerprint computing services.
[90] W3C MAWG = Media Annotation Working Group http://www.w3.org/2008/WebVideo/Annotations/ Joanneum Research is the PrestoPRIME partner on the committee
[91] "Signature" http://www.ina-entreprise.com/to-know-ina/signature.html
[92] D4.3.1 A lightweight audiovisual fingerprint technology (INA, Prototype); PrestoPRIME deliverable
[93] *Competence Centres* have become a feature of European technical support and coordination. PrestoPRIME has detailed descriptions of the intentions for PrestoCentre, available on the project website (the D6 set of documents, http://www.prestoprime.org/project/public.en.html). A general

audiovisual preservation. PrestoCentre will build and maintain registries of various sorts: archives and archive managers, service suppliers, technology suppliers and integrators. With regard to the specific issue of provenance, PrestoPRIME will also demonstrate use of a content registry. The basic technology itself has already been tested on 200 000 hours of content, with statistics on false alarms and misses from detection of 7 million repeats (of small segments, not whole programmes).

Rights Management: there have been many projects about rights management. In PrestoPRIME, we are moving beyond general projects on rights description structures and languages to implementations of right descriptions and processes that actually *define the working practices, within broadcasting, of the PrestoPRIME partner archives*. So PrestoPRIME is not about a standard so much as an actual system, for automation of the rights clearance processes now handled almost exclusively manually by most broadcasters. PrestoPRIME will also implement rights data within a SIP, using the MPEG-22 standard (inside the overall METS container).

.6.5 Processes

PrestoPRIME will collect information on existing technology and develop new technology where needed to bridge the gap between the MXF world and the OAIS world. But technology exists to be used, and so PrestoPRIME will also integrate technology into working systems, in three areas:

- digitisation
- digital preservation
- digital archive management

.6.5.1 Digitisation: automation for quality control

There is a long history of analysis of audio and video to replace human processing. A major application has been in image restoration, where algorithms developed over a span of decades[94] have been used to find potential problems in audio, images and video – and in some cases correct or at least disguise the defects. The technology was further developed for audio by the IRT and Cube-Tech to form the Quadriga[95] audio digitisation workstation, which not only digitises, but also checks the audio signal for areas of potential problems, and produces reports documenting the results. A human operator then only needs to check the noted areas, rather than performing continuous checking, with a large potential increase in throughput. The predecessor of PrestoPRIME – the Presto project which ran from 2000-2002 – was started partly to learn from the Italian broadcaster RAI about their experience digitising 200000 hours of audio in just over three years, using similar automation developed by Reply[96].

PrestoSpace, the successor to Presto, went much further and gathered virtually all the European specialists in video restoration, developing a combined toolkit and

---

review is: Competence Centres: State of the Art Review; Deliverable 5.1: Report on the Design, Value and Impact of Competence Centres
http://www.digitalpreservationeurope.eu/publications/competence_centre_SoAR_1.pdf

[94] Motion Picture Restoration: Digital Algorithms for Artefact Suppression in Degraded Motion Picture Film and Video: 1998 ISBN:3540760407; Anil C. Kokaram; Springer-Verlag London, UK
http://portal.acm.org/citation.cfm?id=522072

[95] http://www.cube-tec.com/quadriga/index.html

[96] http://www.reply.eu/upload/File/cms/content/1825_img_DISR07_capture_audio_eng.pdf-id=1825

architecture. That work was organised by Joanneum Research, who wrote the DIAMANT [97] software that is now the market leader in (digitised) film restoration. PrestoSpace also had as a partner the developer of SAMMA[98], the robotic system for archive digitisation (of cassette formats). SAMMA also depends for its success on signal processing algorithms that detect problems, correct where possible, and provide full XML reports of the results.

In PrestoPRIME, Joanneum Research is concentrating on applying that accumulated knowledge to the specific problem of detecting potential defects in visual signals. Joanneum is working with BBC R&D is integrate the algorithms into a digitisation workflow, matching computer and human tasks to achieve significant increases in throughput by using the same overall philosophy as in the Quadriga product and in the RAI audio digitisation workflow.

.6.5.2 Digital preservation: operating an OAIS

PrestoPRIME will create a digital preservation system, creating a parallel to the commercial system Rosetta developed by ExLibris. PrestoPRIME tools (where relevant) will be used to extend the Rosetta system to support broadcast audiovisual formats. In particular: D1O, uncompressed video and JPEG2000 encoded video (or digitised film) will be supported.

However PrestoPRIME will also integrate all the PrestoPRIME tools into an open-source digital preservation system. The integration will include the testing of a complete digital preservation workflow – and simulation of issues like migration and emulation (using *multivalent*) that are much discussed, but not often tested, and almost never tested on professional broadcast files.

.6.5.3 Managing digital archive services

As archives move from physical items on shelves to files on mass storage, they encounter a range of requirements that they can try to meet with in house skills, but might instead outsource to a service provider:

- The digitisation of the shelf-based content: large archives may have all the equipment and skills needed, but most archives will need help.
- Storage of file-based content: archivists may wish to become IT experts, but many would rather use trustable, cost-effective services.
- Access: the move to file-based content is accompanied by the move to online (web) access. As with mass storage, archivists may wish to acquire web development and hosting expertise, but areas such as *streamed access to video* may well be contracted to specialists.

There is a technical side of outsourcing: setting up of Service Level Agreements (SLAs) to define just what is being purchased – and then being able to monitor (measure, assess) the actual performance of the contractor, and degree of compliance with the SLA. SLAs for storage and for online access services should include access and delivery time specifications that could be – and should be – continuously monitored. Paying for a service that is guaranteed to be available 99% of the time (rather than 97% for a cheaper service) could be a waste of money if the

[97] http://www.hs-art.com/html/products/diamant.html
[98] http://www.fpdigital.com/Products/Migration/Default.aspx

availability is not measured, or even measurable. PrestoPRIME has defined an overall approach to the technical side of such contracts.

.6.6 Tools

Much of the PrestoPRIME toolkit has already been described, because the tools have been part of the work on strategy, use of standards, and implementation of processes already described. This section summarises the full toolset as of the end of 2010.

The tools will all be tested, first stand-alone and then within PrestoPRIME's own architecture for digital preservation. Relevant tools will also be tested on the ExLibris Rosetta system, the digital preservation architecture developed originally as the New Zealand digital preservation system[99] and now a commercial product[100].

Specific tasks of PrestoPRIME software include:

- implement OAIS on broadcast files
- implement preservation strategy according to the PrestoPRIME flow chart
- implement OAI-PMH on legacy metadata
- support time-based annotation
- validate metadata, and map metadata to a common schema
- implement rights management ontology
- perform rights management tasks, automatically
- perform content fingerprinting, and check fingerprints against a database

.6.7 Public Value

The five national audiovisual and broadcast archives that are the main PrestoPRIME partners (BBC, INA, ORF, RAI and Sound and Vision) will establish under the PrestoPRIME project a networked European Audiovisual Preservation Competence Centre, PrestoCentre. This centre will bring together the expertise and experience in AV digitisation and preservation that these archives have developed, through their internal digitisation projects (more than one million hours already digitised by the five institutions) and through their decade of collaboration in the Presto, PrestoSpace and now PrestoPRIME projects.

The PrestoCentre will support saving Europe's audiovisual heritage. The aim is for a permanent body, not just another three year project. The business model is under development, and PrestoPRIME is 'watching with interest' related efforts, such as the recent Open Planets Foundation[101] created at the end of the Planets projects (run by the British Library and concentrating on digital preservation planning and decision making).

The impact of the Competence Centre to be established under PrestoPRIME will be much wider than is generally the case for the dissemination activities of European funded R&D projects. The Competence Centre will not only demonstrate and evaluate the PrestoPRIME project's R&D output through best practice showcasing and training, but will be responsible for reinforcing, leveraging and sustaining at the

[99] http://www.natlib.govt.nz/about-us/current-initiatives/ndha
[100] Rosetta: http://www.exlibrisgroup.com/category/RosettaOverview
[101] http://www.openplanetsfoundation.org.

European level important competences and services for digitisation and digital preservation.

The Competence Centre provides a formal structure for allowing broadcast archives to generate public value by releasing not just their archive content, but also their digitisation, preservation and management knowledge. The knowledge and experience of broadcast archives is just what the other audiovisual institutions need to preserve all that can be preserved of audiovisual heritage – from wax cylinder recordings already in the British Library, to personal recordings on domestic audio cassettes, lying in attics (everywhere), to film and video content in all sorts of non-archive institutions (hospitals, trade unions, industry and commerce – the difficulty would be naming a place that does not hold audiovisual material, all at risk).

.6.8 Steps Toward Creating a Broadcast Digital Archive

1. Broadcast archives need a safe place to keep file-based content.

PrestoPRIME can point broadcast archives to the standards accepted in the general archive, government and national library world for trusted digital repositories[102].

There is a difference between

- a production-oriented system primarily supporting creation of new content;
- a heritage-oriented trusted digital repository (or digital archive).

2. Broadcast archives need to decide the archive format for file-based video content.

In the past, the physical and technical characteristics of what was kept was never defined as the responsibility of the archive: what was kept was determined entirely by the technical standards across production. Now archives have a chance, at least, of specifying or influencing an *archive file format*.

PrestoPRIME can inform this decision. PrestoPRIME explains the problems of certain encodings and file wrappers, and explains the multiple advantages of unencoded video. PrestoPRIME also shows how to estimate costs, over very long periods, of various storage and usage options.

PrestoPRIME explains the significance of failing to archive the best available quality, even if that quality is higher than current production and re-use requirements. Most of broadcasting runs on a very short timescale, responding to immediate needs. One task of an archive is to take a long view, which shows us that today's requirement do not define tomorrow's. The BBC launched high definition the first time in 1936, with the move from 75 to 405 lines. We do not know the specific requirements of

[102] Digital Repository Audit Method Based on Risk Assessment (DRAMBORA) http://www.repositoryaudit.eu/

broadcasting in 2020, but we do know that if we save the best available now, we've done our jobs properly.

For example, an archive could decide that for the remaining years of standard definition production the delivery format will be a file encoded at 50 Mb/s MPEG-II. If a production area has uncompressed material coming in from the camera, PrestoPRIME provides a summary of all the reasons for saving the uncompressed material, and provides tools for computing the cost of that decision.

3. **Broadcast archives needs to decide the archive format(s) for high definition file-based video content.**

As for standard definition, a decision has to be made. Again, over and above the decision about a delivery format, PrestoPRIME provides:

- the reasons for saving whatever is the best quality available across the production chain
- the road map and rationale (temporary archiving) for dealing with compressed material until such time as it is possible to convert to uncompressed
- cost models for evaluating alternatives

4. **Broadcast archives need a preservation factory for the remainder of their tape-based video content.**

PrestoPRIME has looked in detail on the complexities of digital-to-digital transfers from videotape to files, and has a complete analysis and recommended path for the needed workflow.

In particular, the analysis explains why we don't 'save the bits' from D3 tapes but instead convert to SDI, and why we'll do the same for DigiBeta – but it also explains why, for DV formats, there is a genuine choice: save in DV 'native format' or save uncompressed.

PrestoPRIME has looked in detail at a messy issue: divergent embedded metadata. Different file types (AVI, MPEG, MOV, MXF) have completely different ways of embedding metadata. It can all be read, using tools and libraries such as FFMPEG and VLC – but it is not generally possible to put metadata from one file type straight into another file type, because they have their greater or lesser differences in what metadata fields they take, how they are described, what terms they use, and how they 'store the bits'. This all means that metadata has to be mapped to some common schema or standard, to have any success at migration of embedded metadata.

Broadcast archives are committed to external metadata rather than embedded metadata: an overall database that holds the real metadata needed by the broadcaster. This is an admirable and probably inevitable approach, but it does not answer – at all – the issue of lost metadata as files go into and out of various

applications and systems, and into various kinds of conversions – or into a final trusted digital repository.

PrestoPRIME has tools for mapping some kinds of metadata. Broadcast archives need a careful analysis of all workflows, from camera to final file, to uncover, document and solve metadata preservation issues.

5. **Broadcast archives need a preservation roadmap.**

As well as analysing how material flows across broadcast archives, we need to worry about how files move through time: how files are used, stored, re-used, and become obsolete – and then are passed through some form of preservation process so that their useful life can be maintained.

Once digital formats for SD and HD production have been established, PrestoPRIME gives very clear and simple, yet thorough, flow charts for deciding what to do about these files, and when to do it. In particular, **PrestoPRIME shows that compressed SD in standard formats (MOV, AVI, MXF, MPG) can be archived as is, but that it should never be transcoded to another compressed format.** PrestoPRIME also shows the heightened risks associated with archiving of compressed content, plus comprehensive cost models for accurately assessing reduced storage costs vs. increases in system costs (for decoding compressed material), now and in the future.

6. **Broadcast archives need secure storage.**

Broadcast archives need to decide:

- Whether or not it to have a central mass storage system, or a proliferation of more and more (and larger and larger) distributed storage systems in various standards and under various kinds of control.
- Whether to have a specific, designated digital archive that is under archive control as the (eventual) replacement of the kilometres of shelves the archives currently manage – or assume that the storage (centralised or distributed) provided to support digital production is sufficient for all purposes (the disappearance of any identifiable archive in a world where there is just mass storage, managed by engineers and IT staff and not archivists).

PrestoPRIME cannot dictate what constitutes “sufficient for broadcast archive purposes”, but it can say that **storage itself is not a digital library and does not have any semblance of digital preservation functionality**. A conventional library is not just shelves, and a digital library is not just storage. A digital library has to have a set of controlled processes for file acquisition, recognition, verification, metadata extraction, metadata mapping, catalogue development and maintenance, metadata maintenance (authority control and use of power and sophisticated ontology tools), provenance and circulation control.

A digital preservation system needs to either follow the OAIS standard or do something better. It needs to keep content as well-defined objects, whether they are OAIS ‘packages’ or some other construct. This basic object has to be capable of

holding sets of files, following either the METS standard or something equally powerful. The objects need to be identifiable in terms of preservation risk, so that preservation actions such as migration can be implemented on a whole class of such objects. All these processes need to be automated to the largest extent possible, because thousands or ever tens of thousands of files per day cannot be managed by manual methods.

7 Glossary

Term	Definition
AMPAS	Academy of Motion Picture Arts and Sciences
AIP	Archive Information Package; an OAIS object
AVI	A Microsoft audiovisual file format
AXMEDIS	An EC project about rights management technology
cloud (or The Cloud)	services available via internet connection, usually booked and paid for on an 'as required' basis.
codec	Literally: coder-decoder. Generally, a particular way of representing a signal or image by numbers – meaning a way of *encoding* the signal or image (for storage or transmission) and later decoding the numbers to view the image, listen to the audio or view the video ("render the digital object")
competence centre	A body that has expertise in a subject field, such as paper digitisation. PrestoCentre is the audiovisual digitisation and preservation competence centre to be launched by the PrestoPRIME project.
component video signal	Colour video represented as three separate signals
composite video signal	Colour video representation where colour is combined with black and white (luminance) information to make a single signal (carried on one wire, broadcast as one signal)
compression, compressed	Audio, video and images have a certain bandwidth and dynamic range, and so require a certain minimum amount of data to be exactly represented ("captured"). Statistical properties of the resulting numbers can be exploited to reduce ("compress") the overall amount of data. The process can be completely reversible (lossless) or can throw away information considered to be less significant (lossy).
content-based retrieval	Using images to find images, and its parallels in other media
D10	A digital videotape format; the Sony name is IMX; 50 MB/s MPEG 2 in an MXF wrapper; IMX has a videotape equivalent, making this a transitional format – between digital videotape and files.
decode	See *codec*
DIP	Distribution Information Package; an OAIS object
DMI	The BBC Digital Media Initiative
DPX	A standard for holding video or film as "one file per frame" in a folder of such files;
HD	High definition = 1080 lines (generally) though 720p is also used; many other variants exist.
encode	See *codec*
Europeana	The European Digital Library
fingerprinting	In audiovisual processing, a fingerprint is a mathematical description of a piece of content, usually of far lower data rate than the content itself. Fingerprints can thus be stored in a database and used to identify content from the image or signal itself, providing it hasn't been changed too radically.
INGEX	The BBC open-source system for capturing video and saving

	it, uncompressed, in MXF files
Java	An open-source, machine and system independent computer language. Java applications do not run directly on a computer processor, but instead us a *Java virtual machine*, hence giving them independence and a certain potential for greater permanence than for applications developed using ordinary (compiled) languages.
JPEG, JPEG2000	Image coding standard
lossless, lossy	see *compression*
METS	A standard for combining multiple files into one object; it is the main standard supporting creation of the OIAS information packages: AIP, SIP, DIP
mezzanine file format	An encoding which is compressed and so not highest quality, but high enough so that all needed access formats can be produced from the one mezzanine format
MOV	The Apple QuickTime wrapper format
MPEG	Video coding standard
Multivalent	A technology from Liverpool University that uses Java tools to access content (documents, images, audio, video). Format obsolescence (of a codec/wrapper e.g. mpeg1) can be bypassed providing the Multivalent system has a Java decoder/renderer that works on that codec/wrapper.
MXF	Wrapper format used in broadcasting and digital cinema
native format	The numbers resulting from encoding. Applications or processes that can operate on the native format avoid the need to decode and subsequently re-encode, which can be computationally efficient, and can allow an archive to 'save the original'.
OAI	Open Archives Initiative; a document metadata standard
OAI-PMH	The Protocol for Metadata Harvesting (of OAI metadata)
OAIS	Open Archive Information System
ontology	Generally, a system of categories and relationships. In information science, ontologies have a more general set of relationships than for taxonomies (tree structures), and so have more scope for expressing relationships than can be understood to capture the meaning (semantics) associated with the division into categories.
PREMIS	A standard for Preservation Metadata
preservation metadata	Metadata specifically about the preservation needs of a file, e.g. PREMIS
Presto, PrestoSpace	Predecessor projects to PrestoPRIME
provenance	In archives, knowing the history of an object, and all the significant processes or alterations applied to the object
Quicktime	The Apple audiovisual wrapper
Rosetta	The digital preservation system of ExLibris
SD	Standard definition = 625 lines for European television; 525 for North America
SDI	Serial Digital Interface – the standard for digital video
significant properties	The aspects or dimensions or qualities of digital content that

	need to be preserved
SIP	Submission Information Package; an OAIS object
SLA	Service Level Agreement
SMPTE	Society of Motion Picture and Television Engineers – a standards body and professional association
social-networking	Technologies and applications ranging from user-generated tagging and RSS-feeds to Facebook and MySpace, with Twitter and Flickr along the way. All are about interacting with web-content and using web technology to interact with other people.
TCO	Total cost of ownership
TRAC	Criteria for trustworthy repositories http://www.crl.edu/sites/default/files/attachments/pages/trac_0.pdf
transcode	Strictly, computing a new encoding directly from a different, previous encoding *without* an intervening decode back to an unencoded representation. Generally, a decode-recode sequence of computations to change from one coding system to another.
trusted digital repository	A system for storing digital objects that satisfies accepted and standardised definitions for secure operations and long-term existence
uncompressed	See *compression*
URI	Uniform Resource Identifier, such as a URL or URN
VLC	A very widely-used, open-source player for audiovisual files
W3C	The World Wide Web Consortium
watermarking	An alternative to fingerprinting for identification of files. In watermarking, numbers are written into the file in ways that try to be permanent and tamperproof. The numbers can be hidden, or (for images and video) can be visible just as with a conventional watermark on paper. A very obvious logo in a corner of the screen is an obvious watermark, also called a dog = digital onscreen graphic.
wrapper	Audiovisual files are complex: one file can hold video, multiple tracks of audio, subtitles, keyframes, edit and post-production information, time-code, and all sorts of metadata. Various file formats have been developed to hold (*wrap*) all this information in one file. Examples are MOV (Quicktime), AVI (Microsoft) and MXF (SMPTE standard).