ICT-2007-3-231161

Deliverable D7.1.3

Audiovisual Digital Preservation Status Report

Richard Wright BBC

25-01-2010

Document administrative table

Document Identifier	PP_WP7_D7.1.3_Annual_AV_Status_R0 Release 0
Filename	PP_WP7_D7.1.3_Annual_AV_Status_R0_v1.00.doc
Workpackage and Task(s)	WP7 Dissemination and training T1 – Dissemination and publication of results
Authors (company)	Richard Wright BBC
Contributors (company)	
Internal Reviewers (company)	John Zubrzycki BBC; Beth Delaney B&G; Jean-Hugues Chenot, INA
Date	25-01-2010
Status	Delivered
Type	Deliverable
Deliverable Nature	R = Report
Dissemination Level	PU = Public
Planned Deliv. Date	31-12-2009
Actual Deliv. Date	25-01-2010
Abstract	The current status of audiovisual preservation as of January 2010 is described. The previous reports concentrated on digitisation, which remains a significant issue. This report will introduce the new problem of digital preservation (arising from the results of digitisation), summarise the access issues for file-based audiovisual content, and summarise the contributions of PrestoPRIME.

DOCUMENT HISTORY

Release	Date	Reason of change	Status	Distribution
0 v0.01	30.12.2009	First Draft	incomplete	Confidential
0 v0.02	02.01.2010	Second Draft	basically complete	Confidential
0 v0.03	12.01.2010	Final Draft	complete	Confidential
0 v0.04	18.01.2010	Final Draft – re-checked	complete	Confidential
0 v1.00	25.01.2010	Finalised - Published	Delivered	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 fifth in a series of annual reviews of the status of audiovisual preservation in Europe. The previous 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 problems of digitisation were highlighted in previous reports, because that was the focus of PrestoSpace – and the main problem facing the people responsible for audiovisual collections. Now that there has been a significant amount of digitisation, and an equally significant amount of 'born digital' content entering audiovisual collections, there is a new problem to focus on: *digital* audiovisual content, with its risks and preservation needs.

The document has four sections:

1) **A very brief introduction to PrestoPRIME**, needed because the rest of the document refers to areas of the work of PrestoPRIME;

2) **Digitisation**: this remains the biggest problem, as we show that an estimated 80% of audiovisual holdings remain un-digitised. The section summarise the state-of-the-art of technology, current problems, and how PrestoPRIME is building toward sustainable support of digitisation, through the launch of a *networked audiovisual competence centre;*

3) **Digital management and preservation:** the problems of files: how to manage them so they don't get lost or corrupted, or become obsolete and unusable. We also review formal digital preservation technology, including how it applies to audiovisual content, and the role of PrestoPRIME in filling the gaps in current technology. There is one specific section (3.5) on a problem unique to audiovisual content: material that is digital, but *not* in files – and what to do about it;

4) **Access**, which is the goal and payoff of all digitisation, preservation, conservation and archive management activity. There are specific problems – and potentials – for audiovisual content that are not widely found, if at all, in conventional digital libraries with their focus on text (or still images) and documents. Again, state-of-the-art is reviewed and the contribution of PrestoPRIME is described.

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 2010 is described, as 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. The previous reports concentrated on digitisation, which remains a significant issue. This report will introduce the new problem of digital preservation, which arises from the results of digitisation. After presenting the new problem, the report will summarise current state-of-the-art in coping with the problem. **The major conclusion is that there is significant digital library and digital preservation technology for file-based content, but:**

1) **specific tools usually don't work on professional audiovisual files;**
2) **there is very little use of the general technology within broadcasting, though the situation is better in national audiovisual collections.**

Preservation is about access. PrestoSpace always used[4] the CCAAA definition[5] of preservation: *"Preservation is the totality of the steps necessary to ensure the permanent accessibility – forever - of an audiovisual document with the maximum integrity".* This report will summarise the access issues for file-based audiovisual content. **There are three points of significance:**

1) **web technology solves the technical issues that have limited access to audiovisual archives, and digitisation solves the logistical issues;**
2) **formal online access, through *digital libraries,* does not have the tools to support time-based content;**
3) **rights issues remain the major unsolved problem limiting public access to the archives of public service broadcasters and national audiovisual collections.**

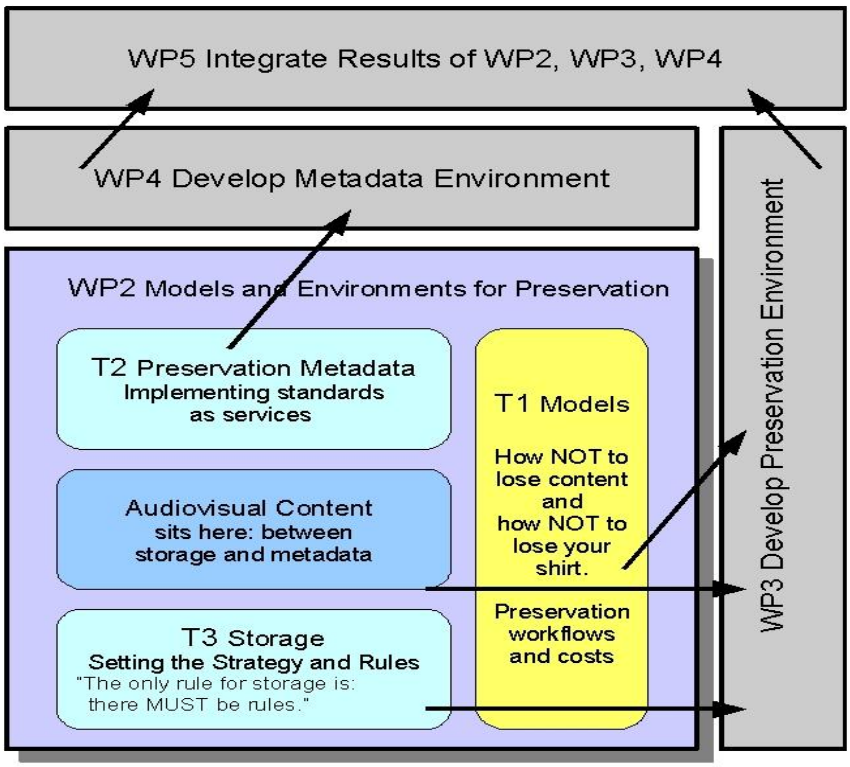
Finally, in each section of this report (digitisation, digital preservation, access) the role of the PrestoPRIME project will be presented.

[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] http://wiki.prestospace.org/pmwiki.php?n=Main.WhatIsPreservation
[5] http://www.ccaaa.org/ccaaa_heritage.pdf

1 A very brief introduction to PrestoPRIME

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.

PrestoPRIME is a large project, divided into various interlocking work packages. Reference will be made to work in these project workpackages and their relation to the information covered in this status report. In order to understand these references, the following figure gives a view of how the technical parts of PrestoPRIME are organised:



This figure refers to four workpackages, numbered 2 to 5. The figure does not do justice to workpackages 3, 4 and 5 – but it shows one essential feature: when audiovisual content consists of files, rather than physical items, it "**sits between storage and metadata**". WP2 will define the basic work of the project; WP3 will implement tools for managing the files – forever! – and WP4 will concentrate on tools for metadata. The results from both WP3 and WP4 are combined in a working, testable system by WP5.

Figure 1: PrestoPRIME technical workpackages

2 Digitisation

Audiovisual archivists have been saying for at least a decade[6] that content on physical carriers needs active intervention because of the threats of format obsolescence, degradation from chemical changes and environmental causes and physical damage to fragile carriers. The Presto survey in 2000-2001 found 5 million hours of content (in just 10 broadcast archives) and estimated that 70% was at immediate risk, while all of it would eventually need some intervention. Presto extrapolated from their finding to estimate the total European audiovisual content, in formal collections, as 50 million hours. Later PrestoSpace[7] and especially TAPE[8] conducted much larger surveys. TAPE found 20 million hours in 400 archives; adding the broadcast archives not included in TAPE gives a total, for Europe, of 30 million hours of content that has actually been formally identified: logged in a European survey. UNESCO estimates total worldwide audiovisual content (in archives or other formal collections) at 200 million hours[9].

There are three basic ways to save this content:

1) preserve the originals
2) make copies of the originals, using the same or similar technology
3) move the *content* onto new technology

Which approach to take depends upon the original carriers. In general, film can be preserved as is (through cold storage), and access copies can be made – also on film. However for audio and video, the technology has moved on: all analogue formats are obsolete (vinyl discs have made something of a comeback, but vinyl is a distribution medium, not an affordable technology for making small numbers of preservation copies). The result is that all audiovisual archives have been digitising their analogue content (or trying to get the funding to do so), beginning as early as the 1980s[10] for audio, and a few years later for video[11].

How much content has been digitised? The Presto (ref 6) and PrestoSpace surveys (ref 7) already mentioned had questions on digitisation projects. Excluding film, Presto found that the broadcast archives were reporting about 60k hours per year of digitisation work, on total holdings (for the 10 Presto-survey broadcast archives) of 4 million hours. That amounts to 1.5% of content being digitised, per year. PrestoSpace had a larger survey, but again mainly of broadcast archives. PrestoSpace found plans for digitisation of 500 000 items over two years, out of a total holdings (within the survey respondents) of 17.5 million items. Again, as a percentage, those figures equal 1.5% of items per year.

Assuming those plans were carried out, and that similar work was carried out in other major collections, and that this work has been going on for a decade, a very optimistic result would be:

Assuming:

- 18 million hours in the major collections[12]
- 1.5% digitised per year (Presto and PrestoSpace result)
- 10 years' of digitisation

[6] At least from the start of the Presto project, in 2000: http://presto.joanneum.ac.at/projects.asp#d2
[7] Final report on users requirements (D2.1) B&G - www.beeldengeluid.nl - Published on 15/09/2004 http://www.prestospace.org/project/deliverables/D2-1_User_Requirements_Final_Report.pdf
[8] *Edwin Klijn and Yola de Lusenet* "Tracking the reel world" 2008 http://www.tape-online.net/survey.html
[9] http://portal.unesco.org/ci/en/ev.php-URL_ID=2034&URL_DO=DO_PRINTPAGE&URL_SECTION=201.html
[10] The first consumer or 'pro-sumer' digital audio formats appeared in the 1980s: audio CD = compact disc, 1982 http://en.wikipedia.org/wiki/Compact_Disc; DAT = digital audio tape, 1987 http://en.wikipedia.org/wiki/Digital_Audio_Tape .
[11] SONY D1, 1986; Ampex D2 1988; Panasonic D3 1991; Sony Digital Betacam 1993 http://www.ultimatewebdesigning.com/articles/formats.html
[12] TAPE survey; 9 million in broadcast archives, 9 more millions in other major archives

Conclusion: 2.7 million hours of digitised audiovisual content

In addition, new content coming into archives (especially broadcast archives) has been digital for the same decade (though one analogue videotape format, BetaSP, is still in common use). Presto results showed that new intake was at about 6% per year for broadcast archives (four times higher than the digitisation rate). Assuming that all new intake was digital, but excluding non-broadcast archives (PrestoPRIME has no data on intake of non-broadcast collections), **the total amount of new digital intake (to broadcast archives) over a decade would be 5.4 million hours.**

Taken together, digitisation and new digital intake total an estimated 8 million hours of digital content, in major collections in Europe.

Eight million hours is a lot, but there are two major reservations:

1) total European audiovisual content in the TAPE survey was 30 million hours. Roughly 25% is digital meaning that **75% is still analogue**
2) the bulk of the digital content is NOT files, but is *non-file-based digital content* residing on digital video tape or DAT or audio CD.

Point 2 deserves elaboration. “Digital” means many things, and is an abused term. Properly, for audiovisual content, *digital* means sound or images that are represented as numbers. Unfortunately many people assume that all digital content is in files, and some even assume that digital means ‘on the web’. Indeed, many people[13] use the verb ‘digitise’ to mean “making content suitable for use on the web” – ie making web-quality proxies.

For such people, almost none of the estimated 8 million hours of digital content would be ‘digitised’ – because in the main it still sits on shelves, on *digital carriers (*Digibeta, DV, M-II, IMX, DV-CAM, Audio CD, DAT, minidisc and more).

The most important conclusion from the above figures is that **a lot of digitisation remains to be done**, because 75% of holdings (in major collections; the figure is likely to be higher in the small collections that hold 40% of overall content) remain analogue, and so at risk to the factors of obsolescence, decay and damage. So while Presto started ten years ago with a problem whose size was about 50 million hours, the size of the remaining problem remains at somewhere above 40 million hours.

Presto and PrestoSpace addressed the digitisation issue. Full information about the nature of the problem, and its solution, are on the PrestoSpace websites: http://wiki.prestospace.org/ http://digitalpreservation.ssl.co.uk/ http://www.prestospace.org/

PrestoPRIME is about the new problem: the 8 million hours that is digital. The problems are:

- how to get it off the shelves and into files: making files out of non-file-based content
- how to get it on the web for access: making web-compatible proxies
- how to preserve the content within those files, forever

The planned contributions of PrestoPRIME to the above questions and to *digital preservation* in general will be taken up in *Section 3 Digital Management and Preservation*, particularly *3.7 Role of PrestoPRIME*. The remainder of this section will look at the state-of-the-art in audiovisual digitisation technology.

2.1 Equipment

Strictly speaking, for both audio and video there are no technical problems with audiovisual *digitisation* technology, if we restrict digitisation to mean the actual conversion of a signal from an analogue to a digital representation. All the problems are playback issues: finding obsolete

[13] This is a personal observation of the author.

equipment, getting it working and keeping it working, and dealing with the consequences of playback faults.

There is a tendency to associate obsolescence with analogue formats. However, because digital audio and video equipment became available in the 1980's, we are now also facing obsolescence issues in the playback of the digital R-DAT audiotape format and early forms of videotape (e.g. D1, D2, D3). The following subsections will describe equipment issue for audio, then film and finally video.

2.1.1 Audio digitisation issues:

Here are the common audio carriers, from the TAPE survey of 4.4 million hours of audio content:

Table 1 - TAPE results on audio formats

Carrier	%
wax cylinders	0.03
coarse groove replicated disks ('78s','shellacs')	1.9
instantaneous disks of any kind	0.2
microgroove disks (LPs)	17.5
open reel magnetic tape	38.6
compact cassettes	20.6
R-DAT	2.5
replicated CDs, DVDs	16.5
recordable and rewritable CDs, DVDs	1.4
minidiscs	0.4
other	0.3
Total	**100**

There is a huge problem reading wax cylinders, but almost no content still resides on that format. The 1.9% of shellac represents 85,000 hours, but most of that can still be read on conventional turntables (with the right arm and needle).

There is a problem with *instantaneous disks* (8800 hours): these were not pressings, they were used to **make** a recording, and consist of a lacquer layer on a substrate. The audio signal could be 'scratched' into the lacquer. Unfortunately the lacquer layer can crack and even peel off. PrestoSpace developed a contactless player[14] for such materials, and a commercial offering was planned for 2009 – but the launch has been delayed.

The most common carriers are LP, open reel magnetic tape, compact cassette and replicated CD/DVD. Playback equipment is available for these formats, though spares and repairs vary by manufacturer and model.

There are two playback issues:

1) deterioration of magnetic tape
2) availability of R-DAT players.

Magnetic tape is a 'sandwich' of various layers and materials, and the active part, containing the magnetic materials, can undergo chemical changes (mainly getting wet and soggy = hydrolysis) which in turn leads to loss of signal, clogging of playback equipment and even complete separation of the magnetic layer from the substrate, destroying the recording.

PrestoSpace did basic research on a contactless magnetic tape player[15], using the magneto-optical principal. That work (and all the PrestoSpace technology) is covered in an INA survey paper[16]. However people generally cope with magnetic tape problems through a combination of heating (baking) and cleaning, reducing the commercial viability of the contactless device.

[14] P4-4 Noncontact Phonographic Disk Digitization Using Structured Color Illumination—*Louis Laborelli, Jean-Hugues Chenot, Alain Perrier*, INA (Institut National de l'Audiovisuel) - Bry sur Marne, France http://www.aes.org/events/122/papers/session.cfm?displayall

[15] P4-7 Improved Magneto-Optical ¼-Inch Audio Tape Player for Preservation—*Marcel Guwang*, Hi-Stor Technologies - Colomiers, France http://www.aes.org/events/122/archiving/session.cfm?code=P4

[16] "Presto – PrestoSpace – PrestoPRIME" Daniel Teruggi, INTERNATIONAL PRESERVATION NEWS No 47 May 2009 www.ifla.org/files/pac/IPN_47_web.pdf

The R-DAT player issue is serious: 112,000 hours of content and the playback equipment has been out of production since 2005. There is conflicting and incomplete information on availability of used equipment and replacement parts, and on refurbishment of parts (notably the head, which wears out just as for any rotating-head equipment). There is also the possibility of playback of audio DAT (R-DAT) on the closely-related data DAT devices, depending upon manufacturer and model, or on home-brew modifications. This paper cannot do justice to this awkward issue, beyond highlighting that the 'R-DAT problem' is a clear case of an issue where somebody should be collecting information, on an international level, and making it available. At the end of this section (see *Section 2.4 Role of PrestoPRIME)* the proposed Audiovisual Competence Centre will be described, including a 'technical watch reports' function which could be used to minimise the effects of the R-DAT equipment problem.

Once the required playback equipment is in place (complete with experienced operators), the remainder of the work in audio digitisation has been well documented by the International Association of Sound and Audiovisual Archives (IASA) in their comprehensive and authoritative Guidelines on the Production and Preservation. of Digital Audio Objects[17].

2.1.2 Film digitisation issues:

While film can last for a century (or more) if kept sufficiently cold and dry[18], there is still a need for film digitisation:

1) making digital copies of deteriorated film elements – or film elements at risk of imminent deterioration through chemical action (colour fade, shrinkage, vinegar syndrome)
2) making access copies
3) preserving film in broadcasting (where film is viewed simply as a carrier, not as an audiovisual medium – because in television all film has to be converted to a video signal before it can be broadcast)

Film scanning equipment is expensive. One of the basic issues in creating the Presto project in 1999 was the high cost of film processes, as compared to videotape. As a rule of thumb, anything involving film would cost roughly ten times as much as a similar operation on videotape. Videotape copying and digitisation had a benchmark cost of €100 to €200 per hour, and film-to-film copying or film scanning/digitisation was indeed running at €1000 to €2000 per hour.

The standard device for making video out of film was the telecine machine, a general family of technology with its origins as old as television itself. Until the 1980's, telecine equipment was analogue: film in, analogue video out. A separate but related range of equipment would scan film at a higher resolution than video. European video standards have 574 on-screen lines, for a picture size of 0.4 megapixels[19] (the quality, however, equates to the image on a 1.6 megapixel camera[20]). The result of higher resolution scanning couldn't be stored on standard videotape, and was stored as data rather than as a video signal, so high-resolution film scanning equipment came to be called *datacine* machines. Telecine equipment was expensive (several hundred thousand Euros, roughly ten times the cost of high-end professional videotape machines), and datacine equipment was

[17] IASA TC-04: Guidelines on the Production and Preservation of Digital Audio Objects, March 2009. Ed. by Kevin Bradley, IASA President and Vice chair of IASA Technical Committee; Printed in Australia, 2009, 150 pp
ISBN 978-91-976192-3-3

[18] Preserve then Show. Ed: Dan Niseen, Lisbeth Richter Larsen, Thomas C. Christensen and Jesper Stub Johnsen
Publisher: Danish Film Institute: 2002 http://conservationresearch.blogspot.com/2008/11/preserve-then-show-2002.html; a 400-yr preservation plan is presented in the article "Environmental Assessment and Condition Survey: A Strategic Preservation Plan for the DFI's Motion Picture Film Collections" Jean-Louis Bigourdan

[19] 576 lines, 702 samples per line = 404352 samples per image; cf ITU-R BT.601 = CCIR 601 = "Rec. 601" http://www.itu.int/ITU-R/index.asp?category=information&rlink=rec-601&lang=en

[20] Stills camera use a Beyer filter to achieve colour from a single sensor

substantially more expensive, up to one million Euros – or more depending upon quality and features.

PrestoSpace worked on technology to produce much cheaper film scanning, aimed at the needs of film archives. With the advent of high-definition video (HD), there is now crossover technology that works at (video) high definition, producing either HD video (typically 1080 scan lines) or 1K or 2K (scan lines) as a data output. The cheaper devices do not work in real time, but the advent of file-based video means that an 'HD video signal' no longer has to be either produced or recorded in real time: it can go into a file, at whatever speed the equipment is capable of producing. The result is a *video file*, and the distinction between telecine and datacine disappears.

The technology supported by PrestoSpace has resulted in a fast, affordable film scanner, with features needed for archive films: sprocket-less handling, very gentle to film (even damaged), robust to archive film impairments (shrinkage, aged tape splices, curling, damaged sprocket holes), brought to market by P+S TECHNIK[21], Munich in October 2009. The basic approach – sprocket-less handling, "flashed" images registered into frames by image processing, data output at 1080 vertical lines or better – is also now being produced by several other manufacturers[22], and prices are down to 1/10 of the cost of the first generation of datacine equipment.

Sound on film: image processing approaches: cinema has included sound since the late 1920's, and has particular ways of recording sound that were not discussed above in the subsection on equipment for sound. There are many variations:

- sound recorded optically or magnetically
- sound recorded on the same carrier as the images, as a 'sound track', or sound on a separate carrier
- optical sound can be recorded as a variation from clear to black over the entire width of the optical sound track, or by varying the width of the clear area (with the rest being solid black).

For optical sound tracks, there is the possibility of scanning (possibly at the same time as the rest of the carrier is being scanned) the sound track, and using image processing to recover the audio. This technology allows the use of some forms of sound restoration that could not otherwise be attempted, such as using geometrical rules to identify non-audio portions of the sound track (e.g. from dirt or scratches). However to the best of the author's knowledge, no image-based system has achieved the dynamic range obtained by "shining a light at it" and measuring the result, in analogue, through a photocell. Analogue optical sound is capable of a signal-to-noise ratio of approaching 60 dB[23].

Over the last decade a range of research groups have looked at optical processing of sound. For this PrestoPRIME review, the author has looked at these projects[24], and it is discouraging to report

[21] SteadyFrame Universal Format Scanner http://www.pstechnik.de/en/scanner_steadyframe.php
[22] Examples: Flashscan http://www.mwa-nova.com/flashscanHD.htm; Golden Eye http://www.iconplus.gr/en/post-production-2a.htm; SCANITY http://www.dft-film.com/scanners/scanity.php
[23] A telecine machine with optical sound and a 57 dB SNR: http://www.ctmdebrie.com/pdf/Telecinemas/p39to42-Telecinemas_HarmonyHD&DixiSD.pdf ; a projector with a 56 dB SNR: http://www.kino-proekt.ru/pdfs/support/e15-head.pdf
[24] 1-Restoration Of Optical Variable Density Sound Tracks On Motion Picture Films By Digital Image Processing Detlef RICHTER et al http://209.85.229.132/search?q=cache:eockZEctjx4J:www.ite.fhwiesbaden.de/~poetsch/download/filmrest_en.pdf+film+optical+sound+dynamic+range&cd=24&hl=en&ct=clnk&gl=u
2-Low-cost and Low-complexity Optical Sound Restoration using Image and Sound Processing Techniques A. Floros, N. Grigoriou, , N. Kanellopoulos, Ionian University, Dept. of Audiovisual Arts, Corfu, Greece http://www.iasa-web.org/downloads/publications/TC03_English.pdf
3-Restoration of movie films by digital image processing Rosenthaler, L.; Gschwind, R. Digital Restoration of Film and Video Archives, IEE Seminar on Digital Restoration of Film and Video Archives London, UK, 16 Jan. 2001 also: Gschwind, R. (2002). Restoration of movie films by Digital Image Processing? In Niseen, D., Larsen, L.R., Christensen, T.C., and Johnsen, J.S. (Eds.) Preserve then Show. Danish Film Institute.

that the basic issue of dynamic range, or (equivalently) signal-to-noise ratio (SNR), is often not mentioned. However the author has heard demonstrations of several of these systems, and has asked the researchers about SNR. From that experience and those discussions, it is obvious (if I can hear the background noise, anyone can) that SNR is at around 40 dB, or less.

A 60dB dynamic range, for sound, requires the ability to measure amplitude over a range of 1000:1, meaning the largest amplitude that can be measured needs to be 1000 times larger than the smallest[25]. This 1000:1 ratio means that the optical sound track would have to be scanned with a resolution of 1000 samples. As we have seen, standard definition telecine equipment scans at about 700 horizontal samples, across the whole image. An HD telecine gives about double that, and datacine typically delivers, at best, 2000 or 4000 lines *across the while image*. The optical sound track is typically less than 10% of the width of the film[26], so scanning at 2K or 4K gives roughly an 'optical image processing' resolution of 200 or 400 samples – which translates into a maximum dynamic range of 26 dB (for 200) and 32 dB (for 400). Until an optical sound track can be scanned with at least 1000 samples just across the sound track itself, no amount of image processing will achieve the basic required dynamic range achievable by a simple lamp and photocell[27].

2.1.3 Video digitisation issues:

As discussed for audio, the actual digitisation of a video signal is not the problem; the difficulty is playback of the old analogue recordings, for a range of reasons:

- lack of equipment, spares and repairs
- lack of experienced operators
- problems with deterioration of older tape, particularly *sticky shed* syndrome[28], or similar issues arising from hydrolysis or other chemical changes
- inability of older videotape players to compensate for degraded signals – excellent time-base correction is built into modern videotape players. Older players require use of an external time-base corrector, which may or may not perform as well as a built-in corrector optimised for that videotape format
- lack of information from the player on signal problems; modern equipment can report line and frame dropout – and "conceal" them if desired (replace them with an adjacent slice of signal)
- various problems associated with how colour signals are handled – this *annual review* can't go into real detail, but colour information is carried in different ways on different formats, and the recovery of the colour can be done in various ways. In particular, colour and luminance ("black and white") information can be carried separately as three *components*, or made into one *composite* signal. Component is always best, as composite requires a

http://www.dfi.dk/NR/rdonlyres/40F7A2C7-6933-4D00-B683-FD7663C28C8C/0/RudolfGschwind.pdf/
4-RESONANCES project http://www.riam-resonances.org/en/
http://cmm.ensmp.fr/~hassaine/restoration.html also- Efficient Restoration Of Variable Area Soundtracks, Image Analysis and Stereology, June 2009 Abdelâali Hassaïne, Etienne Decencière And Bernard Besserer
http://www.wise.com/ias/article.php?id=245&year=2009&issue=6

[25] Sound power increases in proportion to the square of the sound amplitude, so the formula for sound in decibels is 20 times the log of the amplitude. Equivalently, sound power increases by 20 dB for every factor of ten increase in amplitude. 1000 is three factors of ten and so represents a 60 dB dynamic range. http://en.wikipedia.org/wiki/Decibel

[26] An example optical sound track is here: http://history.sandiego.edu/gen/recording/motionpicture.html

[27] A similar calculation can be made in the vertical direction of the sound track, to determine the frequency response upper limit. Sound has to be sampled at N times per second to get a maximum frequency of N/2 (sampling theorem: http://en.wikipedia.org/wiki/Nyquist%E2%80%93Shannon_sampling_theorem). For a 10 kHz upper freq limit, there would have to be 20,000 samples per second – and (modern) film runs at 24 frames per second, meaning 833 samples per frame. For CD quality sound, meaning 44.1k samples per second, there would need to be 1837 samples per frame.

[28] US Library of Congress: Sticky Shed Syndrome in Magnetic Tapes
www.loc.gov/preserv/rt/projects/sticky_shed.html

decode (from the NTSC, PAL or SECAM methods of combining colour with a black and white signal) before digitisation.

Throughout this review, various issues arise that really call for action at a European and global level, and the equipment and skills shortage (in analogue videotape playback) is one. A later section (*2.4 Role of PrestoPRIME: A Networked Audiovisual Preservation Competence Centre*) will discuss how PrestoPRIME is launching a European approach to all these information issues. Various audiovisual archive organisations have discussed collecting information about equipment, and even collecting the equipment itself to refurbish and redistribute where needed. PrestoPRIME will pick up this issue as one of the proposed areas of work of a European Audiovisual Competence Centre.

A "universal videotape reader": out of the frustration with all the various formats and their equipment problems and shortages has come the idea of building a device that would scan videotape (of various widths, presumably) to determine the recorded signal and recover it. This would be a daunting challenge:

- despite 60 years of videotape development, such a device has never been even attempted.
- such patents as exist[29], in the general area of visualisation of imaging of a videotape signal of unknown type, are still restricted to a single physical format.
- scanning speed: (if a magnetic medium can be scanned!) using low speed to achieve a very high-resolution 'image' would imply playback at much less than real time. High-resolution film-scanners struggle to get 4k scanlines on 35mm film, typically scanning a few frames per second = something like a throughput of 20,000 scans per second (5 frames/s at 4K resolution). Standard resolution video (in Europe) has about 700 samples per line, 625 lines per frame, 25 frames per second = 10,937,500 samples. The 10 million then has to be divided to take account of the fact that the signal is not one long line, but a bunch of slanted lines that overlap. The number of parallel lines on a videotape differs according to tape format, head rotation speed and tape recording speed[30]. Taking 100 as a typical value and again invoking the Nyquist Sampling theorem, a minimum of 200,000 scans per second[31] of video signal would be needed – implying a machine that runs 10 times slower than real time, at today's scanning rates.
- scanning resolution: for Betacam[32] tape, one frame of video occupies approximately 115 mm of tape, and holds about 200k samples = roughly 2000 per mm. That's daunting enough, but the frequency has to be detected sufficiently well to equal the signal-to-noise ratio of 'real' videotape equipment. An estimate from PrestoPRIME[33] requires a scanning resolution on the order of 300k scans per mm, about two orders of magnitude beyond the capabilities of current optical and magnetic scanning equipment – and instead requiring the technology such as scanning electron microscopes. It should be cheaper to find a way to make new heads!
- there are many other problems: all the various physical formats, the various analogue encoding techniques, the various signals to be recovered (component of various sorts; composite of various sorts), variation of head alignment (not all heads are perpendicular to the track direction) – and the problem of physical condition of the tape remains, principally sticky shed causing head clog. It's bad enough to clog a conventional videotape machine. One can only speculate on the consequences of sticky-shed on a unique, world-class "scanning magneto-microscopy" universal videotape reader.

[29] US Patent 5046167 - Video tape recorder with a video printing controller – Sony, 1991
http://www.patentstorm.us/patents/5046167/claims.html
[30] http://www.danalee.ca/ttt/video_recording.htm
[31] 10 million times 2 (sampling theorem) and divided by 100 (overlapping lines) = 200 000
[32] http://betacam.palsite.com/format.html
[33] Jean-Hugues Chenot, INA, personal communication. Decoding a 7 MHz signal to 8 bit accuracy requires a scanning resolution on the order of 300k scans per mm: 7 MHz x 256 samples / 50 frames per second = 3.5 million samples. For Betacam with 115 mm per frame, 3.5M/115 = just over 300k.

The requirements for a *universal videotape reader* need to be compared with the technology needed to simply refurbish existing worn-out heads. Companies that advertise such services still exist[34], including making replacement heads. Another role for a European Audiovisual Competence centre is keeping such companies in business, by connecting them to users.

Quality control during transfers

A major issue in digitisation for preservation (as compared with making access copies) is quality control: the new digital content will be the new archive master. While the need to keep originals is recognised, those originals will be of limited value once the machines to play them, plus all the spares and repairs, plus the experienced operators are all gone, a situation only made worse by the chemical changes in the tapes themselves. The BBC response has been to perform manual checking of all the content, using an operator to look at every second of the video, and listen to every second of the audio. This is an expensive process, it takes a lot of time and staff, and there are some errors (e.g. a brief loss of signal such as a line dropout) that humans tend not to pick up.

There is already significant technology for automation of quality control:

1. audio checking built into digitisation systems (Quadriga[35], NOA[36], Reply[37], AudioInspector[38]). These systems, and others, look at physical properties of the audio signal. When combined with manual checking (perhaps through spot checking), the use of the measurements has been shown to be capable of reducing cost while actually increasing quality[39].
2. video checking built into digitisation systems (SAMMA[40]). SAMMA is the world-leading robotic system for audiovisual digitisation, with a range of proprietary technologies (head cleaning, signal monitoring, error logging, time-base correction) implemented in one 'hands-free' system which completely automates the digitisation process. Individual components of the technology can be acquired separately, with SAMMA Solo just implementing the signal checking.
3. image processing technology to detect defects (EC projects Aurora and Brava[41]; hardware Archangel[42]; software DIAMANT[43]; PrestoSpace extended these algorithms, and connected their use to a *preservation workflow* and overall architecture[44])
4. commercial video *file-checking* software, which may include some aspects of audio and image processing (e.g. Tektronix Cerify[45], Interra Baton[46], Harris Videotek QuiC[47], Pixelmetrix VISUALmpeg[48], Telestream FlipScan[49], IneoQuest IQMediaAnalyzer[50])

Beyond these technologies for video checking and analysis, there has been a growth over at least a decade of "workflow tools": systems which organise the use of tools. As broadcasting has sought

[34] http://www.ict-ltd.co.uk/ltd/magnetic-recording-heads.html
[35] http://www.cube-tec.com/quadriga/indepth.php
[36] http://www.noa-audio.com/index.php?option=com_content&task=view&id=22&Itemid=45
[37] http://www.discoveryreply.eu/products/mam/
[38] www.audioinspector.com/
[39] This was the basic result of the mass digitisation of the RAI radio archives, started in the late 1990's. Two descriptions of that work: www.lim.dico.unimi.it/eventi/ctama/RAI.htm presto.joanneum.ac.at/Public/D3_1.pdf
[40] http://www.fpdigital.com/Products/Migration/Default.aspx
[41] http://www.cultivate-int.org/issue2/brava/
[42] http://www.snellgroup.com/products/conversion-and-restoration/restoration/archangel-ph.c
[43] http://www.hs-art.com/
[44] D8.2 Restoration Management Tool http://prestospace.org/project/deliverables/D8.2_public.pdf; D8.3 Audiovisual Defect & Quality Description Schemes and Descriptors http://prestospace.org/project/deliverables/D8-3_RST3.pdf D19.1 Report on project results integration http://prestospace.org/project/deliverables/D19.1_public.pdf
[45] http://www.tek.com/products/video_test/content_verification.html
[46] http://baton.interrasystems.com/overview.php
[47] http://www.harris.com/view_pressrelease.asp?act=lookup&pr_id=1961
[48] http://www.pixelmetrix.com/eng/visualmpeg.htm
[49] http://www.telestream.net/pdfs/datasheets/dat-FlipScan.pdf
[50] http://www.ineoquest.com/iqmediaanalyzer-pro

to reduce staff costs and introduce automation, many suppliers of digital production and playout systems have begun to emphasize workflow. For example (one among dozens if not hundreds): "Pharos[51] delivers improved search, browse and better managed workflows ..." The phrase "broadcast workflow management" gets 11,500 Google hits. As archive content is digitised and becomes file-based, this whole area of technology becomes increasingly important. PrestoPRIME is delivering a report on workflow languages (***D2.2.1 Review of semantic process modelling and workflow languages)*** in early 2010, and has a whole task devoted to workflow: ***WP3T1 Processing and workflows for audiovisual migration*.**

Because of the pressing need to improve the cost-effectiveness of digitisation, PrestoPRIME is looking at automation for checking the signal coming out of a digitisation process. Tools will be developed in the PrestoPRIME work area 3 on **Data management and processing for media preservation**. The existing commercial technology concentrates on *conformance*: whether the data in the file conforms to the rules for that type of file. This is syntactic checking, but not, in general, audio or video defect checking. PrestoPRIME will build on a long history of technology relating to film and video restoration – which did look at the images, not just the file syntax – to develop tools that could make a real difference to the current cost and time bottleneck associated with manual checking.

Moving from digital videotape to files. In this section we have described remaining technical issues around digitisation (moving from analogue media to files), with some indications of relevant work in PrestoPRIME – and of the significant potential of a European Competence Centre. There is a related issue of moving from digital media to files. This is the problem of the audio formats DAT and minidisc, and early digital video formats which are now obsolete (D1, D2 and D3 for starters). Strictly speaking, this problem does not involve digitisation, so it will not be discussed here. Moving from *non-file-based* (though digital) to *file-based* content is a special case of digital preservation, peculiar to audio and video, and with particular problems that are distinct from analogue digitisation, and distinct from the general concerns of digital preservation. This issue is the subject of *Section 3.5: Moving Digital Content into File*.

2.2 Training

There is a need for specialist training in audiovisual issues. The people in charge of audiovisual material have themselves stated (in surveys by PrestoSpace and TAPE in 2006 and 2007) that they frequently had no staff with specialist audiovisual skills. Here are some TAPE[52] results:

> **Do you have staff … professionally trained for working with audiovisual collections?**
> No 217 (61%) Yes 139 (39%)
>
> ***What are the possibilities to be trained for working with audiovisual collections in your country?*** Serious lack 119 (38%) Some, but more training needed 129 (41%) Sufficient opportunities 63 (20%)

TAPE itself ran a series of oversubscribed one-week courses, at venues across Europe and training participants from across Europe. There has not been a successor project or mechanism to again offer training with this European scale and focus. A potential collaboration between Germany, The Netherlands and the UK has been discussed[53], but as of December 2009 funding hasn't been established.

Professional organisations in the audiovisual area do offer training; important sources of training are:

[51] http://www.pharos.tv/
[52] www.tape-online.net/docs/**Tape_survey_factsheet.pdf**
[53] Discussions at the end of the TAPE project between the University of Amsterdam, the HATII group at the University of Glasgow and the Berlin Film Archive (Deutsche Kinemathek).

2.2.1.1 IASA = International Association of Sound and Audiovisual Archives[54]; IASA published the standard guide to audio digitisation and preservation: IASA-TC04 2nd Edition Guidelines on the Production and Preservation of Digital Audio Objects, March 2009. That document is also helpful for digitisation and preservation of video. IASA also organises training courses, or at least has re-activated its training committee (see CCAAA document, just below) and intends to organise more training activities. They sponsored training in Mexico where FIAT (see next) were also a sponsor, and other international activities. Within Europe, IASA has a regional organisation that could be useful is organising training activity.

- FIAT = International Federation of Television Archives. FIAT has been very successful in organising funding for training, and has run courses in South Africa, the former Yugoslavia, Mexico (twice), the Caribbean (twice), Brazil, Chile, Thailand – and others. The past courses are unfortunately not listed in the training section of the FIAT website, but are listing in the 'calendar archive' on the right-hand side of this page: http://www.fiatifta.org/cont/calendar.aspx
- CCAAA = this is an umbrella body of professional audiovisual organisations. They reviewed training in 2003, with a 2006 update: http://www.ccaaa.org/ccaaa_protraindev.doc
- AMIA = Association of Moving Image Archives[55]. AMIA runs an International Outreach Committee, which is perhaps more liaison than training, but it provides essential contacts. A recent AMIA conference session reported on the Outreach work http://archive.witness.org/2009/11/17/reaching-out-at-amia/
- ARSC = Association of Recorded Sound Collections. They maintain a list of training opportunities, mainly focused on North America: http://www.arsc-audio.org/etresources.html
- FOCAL[56] = Foundation for professional training in cinema and audiovisual media; they offer media skills training, aimed at production, which could have some interest for archivists (sufficient for IASA to list them)
- FOCAL[57] = Federation of Commercial Audiovisual Libraries (we know, it is confusing to have two FOCALs; there is another FIAT, as to that). FOCAL has the Jane Mercer fund to support "training, education and promotion of the footage industry" and a training committee to "assist in formulating and monitoring training policy throughout the audio-visual industry". They work with FIAT and would like to work on a concerted approach to audiovisual preservation training[58].

In addition, there is online information from the major library and archive professional bodies.

- IFLA = International Federation of Libraries and Archives has IFLA/PAC[59], the Core Activity on Preservation and Conservation, hosted by the Bibliothèque nationale de France. Their publication International Preservation News had two special issues on audiovisual matters[60] in 2008 and 2009, but they don't actually run training courses.
- ICA = International Council on Archives. This is the major world body for archivists, and has no special section (or training or publication) on audiovisual issues. They do support developing-world archives with an Archival Solidarity section, which maintains a useful database[61] run by Nancy Morelli of Concordia University (Canada).

Individual institutions also provide training, or at least educational activities.

[54] http://www.iasa-web.org/
[55] http://www.amianet.org/index.php
[56] http://www.focal.ch/E/
[57] http://www.focalint.org/
[58] Personal correspondence with Sue Malden, FOCAL chair and FIAT/IFTA programme organiser
[59] http://www.ifla.org/en/pac
[60] http://www.ifla.org/en/publications/international-preservation-news
[61] http://archives3.concordia.ca/Solidarity/default.html

- The British Library has run two Unlocking Audio[62] workshops, in 2007 and 2009; a large international audience attended each of these workshops. The British Library Sound Archive also provides 12-week supported internships[63].
- New York University runs APEX[64], the Audio-Visual Preservation Exchange programme that has organised placements and exchanges with Ghana and Brazil.

Training is available at a national level, particularly in North American and Europe:

- the Conservation Center for Art and Historic Artifacts (CCAHA[65]) offers educational programmes, including a series of 3-day courses specifically on audiovisual content. Unfortunately that series has now ended.
- HATII, University of Glasgow, ran a single one-week course in 2008; http://www.hatii.arts.gla.ac.uk/news/tape.html
- Skillset in the UK is a national training body. In cooperation with FOCAL (the footage FOCAL), they are now (2009-2010) running a half-year course for "ten young people," with two weeks of formal classroom training, and two 10-week placements http://www.focalint.org/focalfunding_aug09.htm
- FOCAL (the footage one) has for many years run an annual one-week course as an introduction to audiovisual content, including components on technology and preservation. They now will run an additional Training Week based in Manchester with Skillset funding, during 8–12 March 2010. http://www.focalint.org/CPDtrainingweek_oct09.htm

There are degree courses in audiovisual archiving:

- The University of East Anglia, UK offers an archive variant[66] of an MA in Film Studies
- University of California, Los Angeles USA. The Department of Film and Television and Department of Information Studies offer the Moving Image Archive Studies[67] program, a two-year Master of Arts degree programme.
- Rochester NY USA; The L. Jeffrey Selznick School of Film Preservation at George Eastman House offers a one-year, international program in Motion picture archival training[68].
- Charles Sturt University, Australia offers a Graduate Certificate in Studies in Audiovisual Archiving[69], available only in distance education mode and delivered online.
- University of Amsterdam, The Netherlands, offers a professional MA (taught in English) in Preservation and Presentation of the Moving Image[70].

Finally, a list of 'resources and advice for screenwriters and film-makers' shows training courses and related activities[71] (mainly in the UK). These are aimed at new production, rather than archiving and preservation.

The above information covers the bulk of the world's training in audiovisual concerns, and was summarised in two pages. While PrestoPRIME hopes the content is useful, the point to be made is that there is very little training, and that is scattered and transient (courses and initiatives come and go). Again, a major role could be played by an *Audiovisual Competence Centre*, to collect such information, keep it up-to-date and put it where people can find it. Above all, there is a need to coordinate all the available resources (from professional associations like IASA, FIAT and FOCAL, national training bodies like Skillset, universities and other formal training institutions, individual

[62] http://www.bl.uk/unlockingaudio
[63] http://www.bl.uk/reshelp/bldept/soundarch/intern/internships.html
[64] http://www.nyu.edu/tisch/preservation/news_2009_2/ghana2009_news.shtml
[65] http://www.ccaha.org
[66] http://www.uea.ac.uk/ftv/Courses/Postgraduate+taught+courses/maarchiving
[67] http://www.mias.ucla.edu/
[68] http://www.eastman.org/16_preserv/16_index.html
[69] http://www.csu.edu.au/courses/postgraduate/audiovisual_archiving_gc/
[70] http://www.studeren.uva.nl/ma_preservation_presentation_moving_image/
[71] http://www.jengovey.co.uk/portal/film_training_schools.html

major institutions such as the British Library, and the various national film archives) to create well-planned, low-cost and above all **frequent** training courses, around Europe and around the world.

2.3 Funding

The Presto and PrestoSpace surveys both found that archives were digitising at a rate of about 1.5% of their holdings, as cited above in *Section 2 Digitisation*. p 7. As anyone can compute, that would translate into 30% of *current* holdings digitised over 20 years, and 45% over 30 years. Audiovisual formats don't last for 30 years, in general, and the tapes themselves may also deteriorate in much less than 30 years, depending upon storage temperature and humidity.

The basic issue is funding. PrestoSpace looked at the above figures and estimated that annual spend, in Europe, on audiovisual digitisation, was €25 million at most[72]. There was a shortfall in funding of at least €35 million – *per year*! Without a *preservation factory* approach the shortfall would be more like €75 million per year. And the result would be loss[73]: 40% to 70% of what archives already have on their shelves, in 2005, would be gone by 2045.

Presto and PrestoSpace did not provide funding. They did do work, in conjunction with other activities, to reduce the cost of archive digitisation – to reduce the shortfall from €75 million to €35 million, per year, across Europe. Where does any of the €35 million come from?

National government interest: the Dutch government formally recognised (Sept 2006) the economic significance of audiovisual media as a component of an information-based economy. This recognition came in terms of hard cash: €173 million in preservation funding, to launch Images for the Future[74]. Some of that has to be repaid directly by archives, through new income coming from new business, but the bulk is expected to be repaid by the general economic boost of having accessible audiovisual content and heritage.

If the other nations in Europe were simply to 'pay their share' in proportion to the funding supplied by the Netherlands (a handy table showing a fair share was published by PrestoSpace[75]), that would be a lump-sum funding[76] of €2.73 billion, which would certainly be most welcome. Assuming the Dutch funding will cover a 10-year project, that same level of funding across all Europe would be €273 million per year. PrestoSpace estimated that €60 million per year for 20 years would be needed, equivalent to €120 million per year for ten years. The Dutch level of investment is more than double the PrestoSpace estimate, but it is covering still images as well, and also funding access projects, not just digitisation. It is in fact encouraging, at least to the author, to see that the Dutch investment and the PrestoSpace estimates are relatively close.

However funding in The Netherlands does not automatically create funding across Europe. There is no simple answer, but certainly all audiovisual collection in Europe (or anywhere) should use the Dutch example. All national governments should know about it, and all audiovisual institutions should know about it. There is a clear task here for a European Competence Centre, to gather details of the *Dutch business case* supporting Images for the Future, and make sure that information is widely and easily available.

European-level interest: 2005 was something of a landmark year. Images for the Future was being developed in The Netherlands (final funding was in September 2006), but at the European

[72] Annual Report on Preservation Issues for European Audiovisual Collections (2004), p10 http://prestospace.org/project/deliverables/D22-4_Report_on_Preservation_Issues_2004.pdf
[73] Annual Report on Preservation Issues for European Audiovisual Collections (2004) p4
[74] http://www.beeldenvoordetoekomst.nl/en
[75] Table 12, p34. Deliverable_22-8_Annual_Report_on_Preservation_Issues_2006 http://prestospace.org/project/deliverables/D22-8.pdf
[76] The GDP of the Netherlands in 2008 was $860K, while for the Euro-zone as a whole the GDP is $13.6M, which is 15.8 times large. The figures given are just multiplying the Images for the Future budget by that ratio. http://siteresources.worldbank.org/DATASTATISTICS/Resources/GDP.pdf

level a concerted policy on digitisation and access was being put together, culminating in the i2010 Digital Libraries Initiative[77].

Within 'book libraries', digitisation is an access issue: getting books online. For audiovisual content, digitisation is a life-and-death issue, but it has the added payoff of producing file-based content which can also go online. The i2010 Digital Libraries Initiative began with a letter (April, 2005) from six major European libraries[78], suggesting the need to form a 'virtual European Library'. The European Commission proceeded to develop a policy (announced[79] on 1 June 2005), and a formal communication "ON THE DIGITISATION AND ONLINE ACCESSIBILITY OF CULTURAL MATERIAL AND DIGITAL PRESERVATION" was issued in late August. That is a remarkably swift development. After years of funding technology research (the IST digital libraries strand), and funding only *coordination* at the European level[80], the EC were throwing their support behind an actual European level 'thing' – a virtual digital library, or even a European answer to Google Books. The exact nature of the 'thing' was uncertain, but the outline was there: a web presence; a single 'place to go' to access European digitised content.

We now (December 2009) know much more about this 'thing': it is called Europeana[81], it exists, and it is very important to all audiovisual collections, because it will be a centre for access, a de facto standard for methods and technology for access, and a centre for a range of projects dealing with digitisation, preservation and online access – and several of these projects are specifically about audiovisual content.

Europeana launched on 20 November 2008 (with a catalogue of 2 million items), and promptly sank under its own success. This hiccup should be seen as a hopeful, because projects that stumble from over-access have a much better prospect for the future than projects which have no IT problems – because they also have no users. Europeana re-launched with a more resilient platform in early 2009, and is expecting grow to 25 million items[82] by early 2012.

The projects associated with Europeana are listed here: www.version1.europeana.eu/web/guest. The list includes the following of particular relevance:

- Europeana Connect adds sound material to Europeana: www.europeanaconnect.eu/
- European Film Gateway (EFG) aggregates cinema related material: www.europeanfilmgateway.eu/
- EUscreen contributes television material to Europeana. EUscreen doesn't have a website, but it is building on the work already done in VideoActive: www.videoactive.eu/
- PrestoPRIME "tackles long-term preservation of digital audiovisual material" – which is what the Europeana website says, and certainly PrestoPRIME is about digital preservation. But we are working with Europeana for **access**: audio and video are not text, and PrestoPRIME will develop, implement and deploy the time-based tools that audiovisual content needs. A full section of this document (*Section 4 Access*) will cover the state-of-the-art of online audiovisual access, and give an outline of the PrestoPRIME work.

Other European support: while it is essential for European audiovisual collections to know about Europeana, and work with it, Europeana is not the end of the story. As has been shown very

[77] http://ec.europa.eu/information_society/activities/digital_libraries/background/index_en.htm
[78] http://ec.europa.eu/information_society/activities/digital_libraries/doc/letter_1/index_en.htm The author thinks all this i2010 activity arose from his lobbying of the IST head of Digital Libraries, during the eCulture conference in Graz in May of 2004, but there were other influences.
[79] http://ec.europa.eu/information_society/eeurope/i2010/key_documents/index_en.htm#i2010_Communication
[80] eg Minerva http://www.minervaeurope.org/
[81] http://version1.europeana.eu/web/europeana-project The actual virtual Library is here: www.**europeana**.eu
[82] Outline Business Plan for Europeana, November 2008, p15. http://version1.europeana.eu/c/document_library/get_file?uuid=0c6c6078-8026-4297-9367-dd6d14b73c2e&groupId=10602

effectively by the Hellenic National Audiovisual Archive (HeNAA)[83], there are preservation funding possibilities as part of general EC support for national initiatives.

The HeNAA is a young institution, founded in 2006. However it arises from 100M€ of funding over the last five years, and another 100M€ is expected in the next five years. According to the First Report[84] on the Network of National Coordination (ATHENA): "Within the period 2003-2007, some 100 M€ were devoted to digitisation activities in Greece. The new Digital Strategy 2008-2013 foresees a similar amount of funding for advancing the area developments."

There are many other ATHENA projects, such as PACKED (Flemish content from Belgium). http://www.packed.be/en/projects/readmore/athena/ and Michael http://www.version1.europeana.eu/web/guest/details-athena/.

These are not research projects – they are development and coordination projects. However for some areas of Europe, digitisation projects (as seen in Greece) can funnel substantial amounts of regional development funding into audiovisual preservation. This is money that would be flowing into these nations or regions anyway – but it needs the efforts of a body like HeNAA to organise proposals and business cases in order to obtain a share of that funding.

Finally, there is research at the European level (beyond PrestoPRIME) that is relevant to audiovisual preservation. Most of it will be covered in *Section 3 Digital Management and Preservation*. However one project is specifically related to digitisation, but in document scanning. Project IMPACT[85] is about support for efficient large-scale scanning, and so is, in its sector, addressing the same issue that Presto faced starting ten years ago: efficient technology and workflow for mass digitisation. But IMPACT is also a 'network of centres of competence'[86]. They aim to "Build a network of competence centres in order to provide a single access point for all players involved in mass-digitisation and full-text generation"[87].

This aims of IMPACT are thus similar to the aims of PrestoPRIME – sustained support, using some sort of 'network of competence centres' (IMPACT), or a Networked Competence Centre (PrestoPRIME). So what is a Competence Centre? The next sub-section explains the PrestoPRIME approach.

2.4 Role of PrestoPRIME: A Networked Audiovisual Preservation Competence Centre

PrestoPRIME is aimed at the problems of audiovisual files, not at analogue content sitting on shelves. However, with only 1.5% of analogue holdings being digitised per year[88], digitisation remains a major issue.

The part of PrestoPRIME that is intended to support digitisation, including support after the end of PrestoPRIME (in mid-2012), is the Competence Centre. This Centre is being developed during the course of PrestoPRIME, and its functions are to be launched in stages during the next two years.

[83] http://www.avarchive.gr/default.php?pname=&la=2 also http://www.minpress.gr/minpress/en/index/other_pages-1/ministry-audiovisual-archive.htm
[84] Marzia Piccininno 30 April 2009 ECP-2005-CULT-038099 ATHENA D5.1 http://www.athenaeurope.org/getFile.php?id=273
[85] http://www.impact-project.eu/
[86] http://www.impact-project.eu/about-the-project/concept/
[87] http://www.impact-project.eu/about-the-project/objectives/
[88] see Section 2.3, first paragraph

2.4.1 Structure of the Competence Centre

Competence Centres are not a PrestoPRIME invention, and indeed there is already considerable background on what such centres are, and do. PrestoPRIME has a full workpackage (WP6) devoted to the Competence Centre.

Competence Centres, under various definitions, have been funded by the EU for a wide spectrum of disciplines, ranging from life sciences to semiconductors, and to those focused especially on digital preservation. Competence Centres have become such a standard approach to improving operations in diverse fields that the EU has funded a project, COMPERA[89] to study best practices in Competence Centre operations. This section reflects the findings of the COMPERA Project, as well as other experiences at networked European competence centres.

More specifically, in its Competence Centres: State of the Art Review (Deliverable 5.1: Report on the Design, Value and Impact of Competence Centres), Digital Preservation Europe (DPE)[90] studied Competence Centres devoted to digital preservation.

By Competence Centre we mean a *networked competence centre*, i.e. a strongly established networking organisation of a limited, core group of AV archives in Europe (initially the PrestoPRIME archives) that will systematically engage in shared research activity and planning, and, at the same time, providing the means to document, repurpose, publish and disseminate the output and experiences to a larger group of stakeholders. They commit to build and address a growing community of AV practitioners (user and peer support group) and will shape and sustain the basic communication platform and registries being developed under PrestoPRIME and continue to offer training and dissemination activities, the extent of which is dependent on the success of a model for sustainability (also developed under PrestoPRIME).

2.4.2 Competence Centre Support for Digitisation

The overall plan for functions and activities of the Audiovisual Competence Centre include:

The Communication Platform: Free Information Web Resources

- Best Practice Guidelines – useful case studies
- Monitoring Reports – 'what's happening' information on projects
- Technology Status Reports – what can be done, and practical uses
- Digitisation and Digital Preservation Planning Tools and Software – continuing development of tools from PrestoSpace
- Business Model Publications – how to identify and estimate costs and benefits
- Promoting Access to Content – how to work with Europeana and other 'aggregators'
- Service Brokering and Vendor Management – understanding how to work with service providers and vendors
- Registries – to enable Service Providers, Experts and Archives to find each other
- Training Materials – updated material from PrestoSpace, plus new material
- News – information on current events taking place, news items
- External Resources – developed by other audiovisual professional organisations and projects

All the above are planned as part of a Competence Centre website to be developed during 2010, for launch at the end of 2010.

For people and organisations formally registering with the Competence Centre (thus becoming members) there will be additional information – generally more detail.

[89] see http://www.competence-research-centres.eu
[90] see http://www.digitalpreservationeurope.eu

Online information is all very well, but it doesn't solve all problems. In the author's experience, most people with responsibility for audiovisual content want a dialogue, not just a place to read pre-packaged information. The Competence Centre wants to provide 'someone to talk to'. PrestoPRIME is working on ideas ranging from brief (but free, or nearly free) 'ask one question' sessions, to full consultations which could even include a visit.

Support for digitisation: many of the items listed above are relevant to digitisation. A lot of relevant information was prepared by Presto and PrestoSpace, and by other projects, activities and organisations around the world – and the **Communication Platform** intends to provide access to that information, including improving the information itself, plus keeping it up-to-date and improving its presentation.

In previous sections of this report, the Competence Centre has been mentioned. In the following list, those 'hints' about the role of a Competence Centre and collected, and more fully explained:

- "'technical watch reports' function which could be used to minimise the effects of the R-DAT equipment problem" (*Section 2.1.1*). The technical watch report would cover availability of playback equipment: where to find suppliers of spares and repairs, and the possibility of using data-DAT equipment on R-DAT tapes.
- "Various audiovisual archive organisations have discussed collecting information about equipment, and even collecting the equipment itself to refurbish and redistribute where needed. PrestoPRIME will pick up this issue as one of the proposed areas of work of a European Audiovisual Competence Centre." (*Section 2.1.3*). A *registry* of available equipment is one possibility, or a *Technology Status Report* on such registries, with recommendations for how to dispose of surplus equipment, and where other people can find such equipment – or a *forum* or *email list* could be useful, or the communication platform could just point people to the AMIA and IASA email lists which are already actively used for equipment finding/disposing issues.
- "Companies that advertise such services still exist, including making replacement heads. Another role for a European Audiovisual Competence centre is keeping such companies in business, by connecting them to users." (*Section 2.1.3*) This role has been clearly defined: keeping a *registry* of service providers, with information of sufficient detail so that providers of 'videotape player head refurbishment' can be found. Additionally such companies can use the Competence Centre to find potential customers – again helping the providers of these specialised services to stay in business.

3 Digital Management and Preservation

> "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."[91]

The preceding section dealt with digitisation. After digitisation, an archive is left with files. New archive content may also arrive as files. A library is not just a heap of books, and a digital collection is not just a heap of files. The files need to be managed, in various ways. That management, including the management processes and technology usually called *digital* preservation, is described in this section, organised as follows:

- Background and Context- and introduction
- Management Technology – a survey of file management tools and systems
- Preservation Technology – technology for keeping content working despite technology change
- Audiovisual Requirements – the needs of time-based media
- Moving Digital Content into Files – the particular digital management issue posed by all the audiovisual content that *is* digital, but not in files (because it's a CD, DAT, Minidisc, DVD or digital videotape).
- Projects – current relevant work
- Role of PrestoPRIME – what PrestoPRIME is doing about audiovisual digital preservation

Material is archived so that it can be kept, for use by future generations. Digital material is kept in files, on some sort of digital storage. Files are invisible, and so are easily lost, unless managed by processes that are much better than standard 'file management systems'. *Digital library technology* introduces controlled systems, ideally with rigorous processes for bringing files into the controlled system, and equally rigorous processes managing the change, deletion, copying and distribution of files.

However digital library technology does not directly address the problem of obsolescence of the content of the files: the way text, still images, sound or moving images are represented by bits. Obsolescence is addressed by *digital preservation technology*, which introduces the requirement for a digital library to store not just files, but additionally all the information needed to keep file content readable and usable (so that the content can be 'rendered' (seen, played, viewed), copied, or migrated). The digital preservation approach may include saving obsolete software, and maintaining emulations of the equally obsolete computer systems needed to run it.

3.1 Background and Context

Audiovisual collections are entering a new world. We have systems and experience for dealing with 'things on shelves' that have been built up over hundreds of years, because standard library and archive practices can be applied to these 'things on shelves'. Now this content is being converted to files, and the basic stark facts about files, from a collection's perspective, are:

- we don't know how to manage files;
- we don't know how to preserve files.

The 'we' in the above refers to everyone, not just audiovisual collection managers. The IT industry can create huge numbers of files, but standard software (for instance, Microsoft's file manager, Explorer) has only the most primitive functionality. Microsoft Explorer cannot add metadata to a file, or build a catalogue of files. Files are found by looking in the right place, and recognising the right name – or by using a very tedious search that again can only look at file names.

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

People getting started in work with files often begin with 'file-naming conventions'. From an information management perspective, this activity is like handing out fans in Hades: it doesn't address the real problem – which is that ordinary IT systems don't provide tools for organising files which have anything approaching the sophistication and functionality of standard library processes: acquisition, classification, cataloguing and circulation control.

The current solutions for managing files come from various kinds of specialist software, described in *Section 3.2 Management Technology*.

Effective tools for management of files keep things from being lost (or should do), but do not solve the other major problem affecting files: having found a file, and "clicked on it", it doesn't work. An error message is displayed, or an application attempts to use the data in the file and fails. The failure could be another error message, or it could be video that is scrambled or frozen, or audio that hangs or distorts, or just isn't there.

The problem of "the file that doesn't work" is a preservation issue. All files rely on a whole range of IT systems: storage management, operating system, 'rendering' applications – often with an additional complication that the needed software has to work in conjunction with web technology (an Internet connection and a web browser with the needed embedded application or plug-in or player). Of course the whole computer set-up could be faulty or incomplete, but when a working, complete IT system fails to open a file, there are two main reasons:

- the file has an error;
- the file is in some way outmoded, and relies on technology that was available, but now isn't.

These are both preservation issues. An effective system for file preservation would:

- check for errors when a file was put into the system (preventing some of the reasons for file errors)
- have technology to prevent, detect and if possible remediate any errors that did creep in
- control changes to the overall system, so needed software would not just disappear
- maintain working software for all files in the system, if possible
- finally, if there was no way to keep using a particular kind of file (the needed applications were hopelessly out-of-date, and could not even be emulated, or run on an emulator of a previous IT platform), then the content would be carefully migrated to a modern file, maintaining as much as possible of its original identity.

The technology for doing some of the above, or even most of it, is covered in *Section 3.3 Preservation Technology.*

3.2 Management Technology

A library is not a heap of books, and an audiovisual collection that has digitised and become files should also not be just a heap of files on a mass-storage system. Because standard computer desktop software has very little to offer beyond sticking files into folders (a hierarchical heap), specialist software has been developed to give people better tools.

The 'lay of the land' will be very briefly reviewed, because there is little that PrestoPRIME can do in this area, and most of the technology about to be described (apart from Media Asset Management systems) deals with all kinds of files, rather than having a specific audiovisual focus.

Role of specialist software: computer systems didn't always have files. When the author started, computers would load a programme (from paper tape!), and run it on data. The data would commonly come from punched cards, or paper or magnetic tape. There wasn't much of this data, and the physical container (stack of punched cards, reel of tape) was the organising and identifying mechanism.

Eventually (in the 1980's) computers acquired discs: first 'floppy' discs capable of storing a few hundred kilobytes, and then spinning discs (hard drives) permanently installed in desktop computers that would store a few megabytes. At this point there could be hundreds of files, and systems for organising files became important. Sets of data acquired names and identifying extensions, from operations systems (e.g. VMS, CP/M, OS-1 and DOS[92]) that could manipulate named units of data.

Operating systems could move files to and from storage, and open them, and show the user lists of files on a particular storage device (*directories*) – and that was about all. The functionality that libraries use to manage huge collections was a completely separate world, because computers initially had only small collections, and people could get by with boxes of floppy disks, and lists of files.

However today typical desktop computers have hundreds of thousands of files[93], equivalent to the number of books in a middle-sized public branch library, and still have no real file-management functionality – in the operating system, e.g. Windows) – for indexing, cataloguing and controlling these files. For those many areas of business activity where some such functionality is essential, a range of applications has been developed, in the following main areas:

- indexing and search
- document management
- asset management
- digital archives
- digital libraries and repositories
- digital preservation

What follows is a brief introduction to this technology.

3.2.1 Indexing and search

An index is a quick way to find things. If every item in your house had a label tied to it, and you had a notebook giving the location of every item, nothing should get lost. If the labels used a small number of standard terms, the notebook could be alphabetically arranged, and finding a lost torch (flashlight) would be a matter of knowing the right index term (for instance, anything making light might be called a *lamp*, a preferred term) and then turning to *lamp* and seeing a list of all the lamps in the house, including the desired torch.

Libraries and computers and files are commonly full of text (data that is interpreted as words), which offers an automated substitute for indexing, which unfortunately is now also called indexing. The substitute is to make a notebook showing the location of every word in the system (file, storage unit, archive, entire Internet). A *full-text search* can then be used to attempt to track down the desired file. Full-text indexing isn't as powerful as manual indexing using a controlled vocabulary, but it can be automated and so it does scale to very large tasks, such as finding text in the Internet.

Standard word-processing applications now support some of the functionality needed for indexing. For instance, I'm typing this document in Microsoft Word, which supports a *properties* function for each document, and one of the properties is keywords. This makes the document look like it could be manually indexed, though really this functionality is of little use, for several reasons:

- the keywords are available within Word, but not easily available from other applications. For instance, the *properties* function (same name, different function!) for this very file, using Explorer, shows nothing about these keywords, or most of the other internal properties.
- there is no controlled vocabulary

[92] http://www.osdata.com/holistic/age/age.htm
[93] http://www.dslreports.com/forum/r19536152-How-many-files-on-your-computer

- there is no control of use of the vocabulary (such as a *pick list*), so I could easily mistype things
- most importantly, there is no way to search on the keywords, at least not by standard Microsoft Office applications!

What is possible, in standard Microsoft and Apple applications, and in products from other vendors, is support for full-text searching[94]. It is accessible from Explorer – as the 'Indexing Service' option.

Other applications that allow all text to be word-level indexed for support of full-text searching include:

- Copernic[95], a commercial product for desktop computers
- Lucene[96], an open-source Java application which is generally applied in a web environment
- Autonomy[97], a commercial application at the enterprise level, which also has *artificial intelligence* functionality to aid search and retrieval

A document collection that had been digitised to files could use one of the above approaches to support full-text searching for the entire collection. The resultant functionality would add a fourth aspect to the management of the files:

- organising the files into folders;
- using an orderly naming convention;
- maintaining some sort of list of the files (as a document or spreadsheet);
- full-text search as a finding aid.

Unfortunately, audio and video files are not text. Even more unfortunately, the text that is in these files (the metadata) is NOT accessed by any of the standard full-text indexing tools. The result is that audiovisual collections really need to move up to *asset management* or *digital library/repository* software to get any improvement on simple filename conventions, orderly file structures and basic lists.

3.2.2 Document management

Full-text search does nothing for actually controlling a collection of files, to provide something like acquisition, indexing and circulation control. In the world of electronic documents, the shortcomings of standard computer tools and the need for electronic documents that could be trusted (for legal and other purposes) has led to an entire industry of document management systems. As with full-text search, these systems offer little or nothing to support audiovisual files – but they do show how files can and should be effectively managed.

Electronic document management is now a mature industry: the provenance of a file can be controlled to legal standards, changes can be logged (and made reversible), multi-level access control is possible, and documents can be tagged or indexed (including through use of a controlled vocabulary).

All this is very attractive to all digital collections, including audio and video. Unfortunately, as with applications that perform text indexing, document management systems do not generally support standard audiovisual file formats. When applications that started as document management systems do begin to support audiovisual content, they tend to re-brand as *enterprise content*

[94] This functionality is "turned off" on the standard computer configuration where this author works, but I'm assured that word-level indexing does work within Windows.
[95] http://www.copernic.com/
[96] http://lucene.apache.org/
[97] http://www.autonomy.com/content/Technology/evolution/evolution-of-search-pan-enterprise-search/index.en.html

management systems (as with Documentum[98], which now has products for *records management*, for *web content management* and for *digital asset management*).

3.2.3 Asset management

Digital asset management (DAM) and media asset management (MAM) systems begin to offer the functionality that audiovisual collections need. The area of asset management systems includes the products developed in the last 15 years that support audio and video (and image) files. Asset management as a label is larger than audiovisual content, so not all asset management systems support audiovisual content, but instead might concentrate on physical assets (inventories) or intellectual assets (important for tax write-off purposes).

Asset management systems that do support audiovisual content have two areas of important functionality:

- **metadata** – these applications do (generally) read and write the metadata parts of audiovisual files, and so have the basis for making a proper catalogue, and for full-text indexing of the textual parts of these files. Some can use controlled vocabularies, and some conform to international standards for audiovisual metadata – though the present diversity of standards in that area remains a problem.
- **manipulation** – these applications are very strong on the user interface: the tools that allow audio and video to be seen on a time line or as a story board, with time-code, and with functionality to point to specific points within the file and create extracts (clips) or metadata pointing to clips (edit decision lists). Such functionality is the 'bread and butter' of audiovisual production (and the archives that support content re-use). A major difference between asset management systems and digital libraries/repositories is the near total lack of audiovisual manipulation tools in the library/repository systems.

Some major examples of asset management systems in the audiovisual fields are:

Artesia	www.artesia.com.
Blue Order	www.blue-order.com
MediaBeacon	www.mediabeacon.com
North Plains	www.northplains.com
Virage	www.virage.com

Applications coming from the broadcast sector that have developed into asset management systems include:

Ardendo (Vizrt)	http://www.vizrt.com/products/#MediaAssetManagementMAM
Dalet	www.**dalet**.com
Front Porch DIVA	http://www.fpdigital.com/Customers/AssetManagement.aspx
Harris Invenio	http://www.broadcast.harris.com/productsandsolutions/DigitalAssetManagement/Invenio.asp

For interest, the UK government has a list of preferred suppliers of broadcast asset management systems[99] – five in total, and three are unknown to this author, showing (perhaps) that this is a developing area with many sources of relevant technology.

3.2.4 Digital archives

There are basically two kinds of digital archive:

- off-line storage: a place to put files that do not need fast access; example: email archive
- an electronic version of a shelf-based archive

Most IT companies use the first meaning, and most audiovisual collections use the second! The first is a place for data that is no longer serving a primary purpose. Today's email is important,

[98] http://www.emc.com/products/category/content-management.htm
[99] http://coi.gov.uk/suppliers.php?page=95

yesterday's (or last year's) email is relegated to the archive. The second is itself a primary purpose: the electronic version of a collection or institution that holds valuable content.

Archives (the institutions, not the IT secondary systems) are the original *aggregators*. They collect content, keep it, and make it available. In so doing, they *create access*, by pulling content into a recognised institution, where it can be sorted, labelled and made available to researchers or the public. The secondary systems used in the computer world *reduce access*: sending data somewhere where it isn't actually lost, but where it does need to go through a *restore* process (which takes time, and could require manual intervention) in order to again be accessible.

We should perhaps speak of primary archives and secondary archives. The content of archive institutions is *primary*: this content represents the primary purpose of the archive. The electronic version – the digital archive – of such content is also a primary archive.

The data removed from spinning discs, such as old email, is a secondary archive. The primary function of the email system is to give access to current or recent data, and older data is put somewhere else, with reduced access because of its secondary importance.

The distinction between these two uses of the word archive is important, because computer systems that were developed as secondary archives are now starting to offer themselves as IT solutions for primary archives, which can lead to a great muddle. For audiovisual collections, or for any *primary content*, the vital issue is to use computer systems which *create* access, not those which *reduce* it.

Example of companies which offered secondary archives, but now also offer primary archive services, are Atempo[100] and Front Porch[101]. They are not mentioned here as any form of official recommendation, but simply as examples – so that this review can be specific rather than just dealing in generalities. Atempo has gone through four stages:

- offering secondary archiving services to the general data industry;
- working with audiovisual media companies who also needed secondary archiving, to deal with their limited capacity to 'keep everything online';
- developing primary archive services for these media industries;
- relabeling those primary services as *asset management* systems, to reduce confusion.

The Atempo and Front Porch examples are instructive: there are multiple kinds of archive systems, and multiple kinds of asset management systems, and individual companies (and individual applications) can develop from one into another. The result could be general confusion: what is asset management (or content management), what is a digital archive?

In the author's view, the answer to this confusion is not found in definitions. PrestoPRIME could produce definitions, but that would change nothing. The answer is in functionality: ignore what the system is called, and ask what it does. Secondary archiving systems reduce access – putting content 'somewhere else'. Primary archiving systems (and asset management systems) treat the archive content as the primary content, and concentrate on tools for dealing with that content. Both primary digital archive systems and asset management systems would have metadata tools. One would expect an archive system to excel in metadata tools, and offer professional library tools. An asset management system would be likely to be weak on the librarianship side, but strong on the media manipulation tools.

A guide to the perplexed:

- support for production – the use of the content – requires manipulation tools (to see a storyboard, extract a clip, make an edit decision list)

[100] www.atempo.com
[101] DivArchive http://www.fpdigital.com/Products/Content/Default.aspx?mrsc=DIVArchive

- support for research may have less need for manipulation tools, but greater need for metadata and librarianship tools (formal indexing, controlled vocabulary, acquisition, cataloguing and circulation control)

It would appear from the above that there is a two-way choice: asset management systems (for the hands-on tools) vs. digital library systems (for the metadata tools). What about digital archives? It would be simpler to ignore digital archives, but they exist and so cannot be ignored. They exist for two reasons:

- the existence of secondary data in the computer world, creating the need for primary data to be 'archived' by (secondary) digital archiving applications
- the existence of shelf-based archives (of many sorts) in the real world, and their adoption of digitisation and web technology to create digital archives

For the purposed of this review and status report, the needs of audiovisual collections can be met by the functionality that should come with asset management systems or digital library systems. This review had to include digital archive systems because:

- 'they are there';
- secondary archive systems are **not** what audiovisual collections need;
- there are a lot of secondary archive systems that need to be avoided;
- there are a lot of IT people who only know about secondary archive systems (and so never use the word 'secondary' – they just call them archive systems);
- there is very useful technology coming from companies like Atempo and Front Porch that began in secondary archiving but now offer tools from both the asset management and digital library worlds.

When shelf-based archives 'go digital' they commonly need functionality that cannot be supplied by one vendor, much less one application. These *institutional digital archives* are major IT development and integration projects. As an example, the digital archive for Washington state in the USA (mainly an e-document archive, rather than audiovisual) lists nine separate commercial vendors/suppliers as constituents (partners) in their solution[102].

3.2.5 Digital library and repository technology

The story so far:

- files need management; file name conventions and folder structures are just the beginning;
- text files can benefit from full-text search;
- real control of text files requires a document management system; but those systems don't work (generally) on audiovisual files;
- media asset management systems (DAM, MAM) are used to hold collections of audiovisual files; they are good at manipulating audiovisual media, and can be weak at library functionality;
- digital archives need to be understood, if only to avoid 'archive systems' which only reduce access (taking content offline) rather than enhancing access through added library/archive functionality;
- full control of a large collection of content requires the librarianship tools (which may be labelled as archive or repository systems) of digital library technology

The digital library approach to dealing with files has well-developed methodologies for the basic tasks of:

- identification/characterisation: what kind of file is it?
- verification: does the file actually conform to the required specification?
- association with rendering technology (software needed to display the file contents)
- quality checking: are there errors or problems with the file?

[102] http://www.digitalarchives.wa.gov/Content.aspx?txt=partners

- migration: moving content from old formats to new
- copying: making clones for protection against loss
- making derivative versions, or proxies: encoding, compressing

There are mature services for digital library processes[103]. JHOVE[104], the Harvard Object Validation Environment, is an open source Java application developed to identify and verify the formats of the seven million objects in the Harvard Digital Library. DROID[105] is a more recent identifier, designed to work with the PRONOM[106] format registry which then specifies which service to call for verification (e.g. JHOVE). PRONOM itself is not a tool, but a registry of file types, and all the associated tools/services needed to carry out standard processes (identification, verification, rendering and so forth). These tools mainly work on documents and still images; at best there is' limited support for time-based content, at worst the tools simply do not work on audiovisual files

Important metadata is inevitably embedded within files, and needs to be brought out for identification and further processing, and for collection/consolidation separately from the files themselves, as a catalogue. In 2007, the National Library of New Zealand Library released a metadata extractor tool specifically for digital library use[107]. All online documents that conform to OAI can be searched for by web-spider technology, using the OAI-PMH protocol for metadata harvesting[108]. Related harvesting technologies allows building up directories / catalogues for methods including RDF and MPEG-7 harvesting, but none are as well-developed as is OAI-PMH.

A digital library (or archive) will need to perform the following operations (at least):

- Acquisition:
  - For new material: bring files into the digital archive
  - Legacy material: digitisation from physical items to files
- Documentation:
  - An archive travels on its catalogue. As archives ‘go digital’, the catalogue becomes the major *value-added service* of the archive.
- Viewing:
  - The archive will have to support a multiplicity of ‘proxies’, because bandwidth will be insufficient to move high-resolution video files as quickly as MPEG-4 (or whatever) viewing files
  - Catalogue search, viewing and rough edit will, ideally, be combined in a single asset-management application
- Re-Use
  - Full-quality material will have to be delivered, as files, to edit suites or wherever else they are needed.
- Asset Management and Life-cycle management
  - There is a set of *birth to death* processes here, based on processes established in the document management world (where they started ‘going digital’ 20 years ago). Principal issues include access control, version control and digital rights management.

The functionality just listed is common to library / archive systems in general. They all have modules for acquisition, cataloguing and circulation control. There are two basic differences between a conventional library IT system and a digital repository:

- **a repository holds the content**, not just the catalogue and support for acquisition, circulation control and other processes;

[103] See http://twiki.dcc.rl.ac.uk/bin/view/Main/DevelopmentToolList for a general list of digital curation tools
[104] http://hul.harvard.edu/jhove/
[105] http://droid.sourceforge.net/wiki/index.php/Introduction
[106] http://www.nationalarchives.gov.uk/pronom
[107] http://meta-extractor.sourceforge.net/
[108] http://www.openarchives.org/OAI/openarchivesprotocol.html

- a repository **prevents loss of content**, or at the least tries very hard to prevent loss – by incorporating processes and technology specifically aimed at insuring the continued viability (persistence and currency) of the content.

Today, most broadcast archives are moving from tapes on shelves to file-based content, and using asset management technology to manage these files. In general this means a reduction in librarianship tools, at least for those archives that had library-type acquisition, classification, cataloguing and control processes.

All these areas tend to be weaker in asset management systems:

- files can get in and out without going through a formal acquisition or control process;
- files don't have to be classified using a controlled vocabulary or a hierarchical classification system; they may get *tagged* using uncontrolled vocabularies (and the errors of unrestricted text) or there may well be no formal indexing
- file content doesn't have to have *cataloguing:* an analytical description of the contents.

However there are full-feature asset management systems that do support library functionality, and when they do, they provide features that standard digital libraries do not provide: metadata (keywords, analytical descriptions) that is tied to specific shots (segments, clips) within a video asset, or tied to specific time points in an audio stream. The whole area of time-based tools is what distinguishes media asset management systems from digital library systems. Time-based tools are essential for time-based media. Media asset management systems have such tools, and digital libraries don't – in general.

This situation leaves audiovisual collection managers in a difficulty: they can have the high-level metadata and overall acquisition and circulation control tools they need, or the time-based media manipulation tools – but not both!

But there is a further area of technology that is not addressed by asset management or digital library systems: *digital preservation*, discussed in the next section.

3.3 Preservation Technology

Digital library tools provide management (so files can be accessed and don't get lost), but do not cover preservation. Files face a range of obsolescence issues, addressed by digital preservation technology:

- methods for ensuring that obsolete files can migrate to new standards and formats; PREMIS metadata, JHOVE and DROID file identification tools, databases of information on file formats (PRONOM, Library of Congress)
- methods for emulating old IT environments to extend the lifetime of obsolete formats; project SHAMAN and the Multivalent approach
- criteria for evaluating the reliability of a digital repository; DRAMBORA
- and finally an overall methodology: OAIS.

A brief (only two pages) but information-packed review of the digital preservation technology and its application to audiovisual content is available from the DPE[109] website (in eight languages); the English version is given in Annex I, below. All the above-named tools and projects are described in more detail below.

Audiovisual collections have difficulty finding anyone on their IT staff who has even heard of OAIS, which rather limits support for funding and implementation. Until the various projects and initiatives just listed (PREMIS, SHAMAN, DRAMBORA, OAIS and others) develop software that enters the

[109] Preservation of Digital Audiovisual Content http://www.digitalpreservationeurope.eu/publications/briefs/

commercial world understood by standard IT staff, implementation of digital preservation technology will largely be limited to national libraries and other major libraries. These institutions big enough to have their own IT staff, recruited and trained specifically to implement technology needed by libraries. The rest of us (broadcasting, media production and archiving, small collections without dedicated IT staff) will find it difficult to implement any formal digital preservation technology. A major role of PrestoPRIME is to provide information to ease that difficulty

In addition to the tools, process test beds have been established by digital library projects (including the FP6 Planets project headed by the British Library, ref 114) which allow digital preservation tools and processes to be tested before turning them loose on actual digital library content.

The outstanding research need addressed by PrestoPRIME is to extend digital preservation technology so that professional broadcast file formats are fully supported. At present there is a large gap between the community that understands OAIS (a major digital preservation standard), and the community that understands MXF[110] (a major professional audiovisual wrapper format).

3.4 Audiovisual Requirements

Because of the two worlds problem, professional broadcast formats (MXF in particular) are unsupported by many digital library and preservation tools. Other 'standard' formats are better supported, but many (e.g. AVI, WMV) are proprietary, which is in itself a preservation problem.

The remaining problems relate to the actual content of the files.

- most AV files are compressed. Whatever 'original quality' was lost in compression, will remain lost. Preservation should maximise retention of quality, a capability that needs to be defined and added to current technology.
- time-based content needs tools with a time dimension (cataloguing, navigation, edit)
- the files are complex. Indeed the concept of a wrapper was developed to recognise the complexity of a typical AV file: multiple signals, multiple kinds of metadata – including time-domain (subtitles) and numerical (time code)
- audiovisual preservation involves many related files: lossless and lossy encodings, multiple proxies (supporting access in multiple formats e.g. Real, Windows Media, MPEG, AVI, Quicktime, Flash), various stages of edit and recombination, and a range of rights information: multiple interested parties, multiple collection agencies, non-uniformity from country to country. A complex of information representing signal, metadata and rights must be preserved.

3.4.1 Audiovisual files

Any system for the long-term preservation of audiovisual content will need to deal with

- the data structure and metadata elements of material stored in the archive
- the dynamic process of preservation itself, which includes selecting media for preservation (e.g. risk analysis), planning the preservation actions, processing (e.g. migration), verifying (e.g. quality control) and subsequently updating the archive with the preserved media.

One of standards for metadata for long-term preservation is PREMIS[111], a data dictionary as well as a data model. PREMIS defines *preservation metadata,* which basically is supposed to be all the information needed about a file, to make sure it can continue to be used into the indefinite future. The metadata area of PrestoPRIME (WP4) is responsible for looking at the relationship between audiovisual content and the PREMIS standard, but WP2 has already made a preliminary analysis of *metadata needed for preservation.* The result is in *Section 3* of *Preservation metadata models*

[110] http://www.digitalpreservation.gov/formats/fdd/fdd000013.shtml
[111] http://www.loc.gov/standards/premis/

and extensions of deliverable *D2.1.1 Audiovisual preservation strategies, data models and value-chains* (to be published in early 2010).

The conclusion with regard to audiovisual files is that the simplest way to preserve digital audiovisual content is by use of **uncompressed data, fully described by technical metadata**. Preservation metadata as contained in the PREMIS standard gives a structure for defining the whole IT environment needed by a particular file type. For uncompressed data, any IT environment will do; any generic player of audio and video will do; and the signal can be easily moved, preserving all the bits as in the original file, from one generic wrapper to another. In short, uncompressed audiovisual data short-circuits the need for most of the complexities addressed by PREMIS, or by OAIS itself.

Quoting PrestoPRIME[112]:

> Problems arise because of complexity. Many encodings can share a common wrapper, so that, for instance, a ".wav" file can contain many different ways to represent an audio signal, ranging from non-linear allocation of bits in samples to highly-compressed data. The situation for video is more complicated just because there are so many file and wrapper formats, as well as so many encoding possibilities. Further, two files made using the same encoder, and wrapped in the same type of wrapper, can still differ enormously. They could differ in their *compression parameters,* so that one MPEG-2 file (for instance) could be broadcast production quality at 50 Mb/s, while another could be completely unsuitable for professional editing, being much lower quality and not allowing edit at specific frames (because the compression averages across a *group of frames*[113]).
>
> This whole situation is very unsatisfactory for long-term preservation, because key knowledge is 'embedded' in players, rather than being captured in formal metadata. Hence the survivability of the content is dependent upon the survivability of the players. Audiovisual content is not unique in this respect – text files are equally dependent upon software that can 'render' their contents. However there are ways to reduce the dependence:
>
> - **better metadata**: successful rendering software can determine what kind of data it is dealing with, by reading and interpreting meta-information from the file; this information could in principle be 'pulled out' of the file and made explicit as formal technical metadata. The audiovisual industry would benefit from much more agreement on where and how to place metadata in proprietary file types, and on ensuring that *all* the decode parameters were part of that metadata.
> - **simpler files**: most of the complexities of audiovisual content are to do with compression methods and interpretation of compressed data. Uncompressed audio is virtually self-describing (or needs no description, beyond: the following is a sequence of audio samples – just work out three parameters and it can play perfectly). Uncompressed video is more complicated but it also is 'just a sequence of samples'. Virtually nothing general can be said about the data in a compressed file, and attempting the playback of a compressed file of an unknown type could well prove futile.

Migration and Emulation: the Planets[114] project is currently developing a set of services, tools, methodologies and frameworks supporting long term preservation planning and management. The project is not looking at audiovisual files, and is concentrating on migration: moving from an obsolete format to a new format. In text files (which can be surprisingly complicated: not just all the 'mark-up features' about different kinds of text and different structural sections, but also figures, tables – and a very problematic area of embedded objects.

[112] D2.1.1 (Section 3.2: Role of technical metadata in preservation)
[113] http://en.wikipedia.org/wiki/Group_of_pictures
[114] http://www.planets-project.eu/

For migration of such complex objects from something like Word to PDF, or from PDF to ODF, something might get lost. Planets has a procedure for trying to minimise loss. The *significant properties* are defined, a test conversion can be run (on a test bed), and the results can be automatically analysed to determine a performance score. Various migration options can thus be compared, and the best chosen.

The alternative to migration is emulation: a system capable of running the 'obsolete' software that is needed to use the 'obsolete' file format. The most recent, and comprehensive, project developing an alternative to Migration is SHAMAN, described below (*Section 3.4.3 Data Management*, along with details about PLANETS).

Risk and Quality Assurance: the DRAMBORA[115] project (which is linked to SHAMAN) provides a way to assess risks to content held within a formal repository. While not specifically looking at audiovisual content, the ideas are general and important. The project says:

> *Within DRAMBORA, digital curation is characterised as a risk-management activity; the job of digital curator is to rationalise the uncertainties and threats that inhibit efforts to maintain digital object authenticity and understandability, transforming them into manageable risks.*

Archives are faced with many challenges, and the temptation is to take the resources available and 'do the best we can'. The real importance of DRAMBORA is to show two things:

- the importance of risk: decay, obsolescence, migration, storage: all the problems can be reduced to risk, and to 'cost of risk'[116]. Looking at problems this way provides objective, quantitative data on which to base management decisions.
- risk management: there are formal ways to model and assess risk, DRAMBORA being one that is designed specifically for digital curation.

3.4.2 Metadata

There is a general taxonomy of digital library and digital preservation metadata:

- descriptive metadata
- technical metadata
- administrative metadata
- preservation metadata

Metadata needs of audiovisual content: many different metadata models and formats exist for describing cultural heritage assets[117]. The temporal dimension of media items is a main issue to be addressed to establish interoperability with other cultural heritage collections. The current European Digital Library metadata model lacks support for representing temporal segments of content and annotating them with specific metadata, which is a common requirement for audiovisual content archives[118]. A data model for audiovisual content that does not have a time dimension creates huge problems: an entire file needs to be accessed just to get to one desired segment; text material can be quoted by page or even line number, but audiovisual content would not have similar *granularity*, making it impossible to reference audiovisual content in an effective manner; content cannot be annotated against time code, making detailed cataloguing impossible.

Although the library community has done significant work to establish a common metadata format, the EDL and the audiovisual archive community have still not achieved interoperability and the efforts for establishing protocols and formats for interchange are at a very early stage[119]. Work

[115] http://www.repositoryaudit.eu/

[116] R Wright, M Addis, A Miller, "The Significance of Storage in the "*Cost of Risk*" of Digital Preservation" The International Journal of Digital Curation, Vol 4, No 3 (2009) http://www.ijdc.net/index.php/ijdc/article/view/138

[117] J. Oomen, H. Smulders, "First Analysis of Metadata in the Cultural Heritage Domain", MultiMatch D2.1, Oct. 2006.

[118] B. Delaney and B. Hoomans, "User Requirements Final Report - Preservation and Digitisation Plans: Overview and Analysis", PrestoSpace D2.1, Sept. 2004.

continues between the EDL project, VideoActive and DISMARC[120], which also intends to develop an application profile for audio objects (using metadata terms from the Dublin Core Metadata Initiative plus the Dublin Core Libraries Application Profile – DC-Lib).

The EC working group on digital library interoperability[121] proposes the use of domain specific application profiles and the establishment of a metadata registry for EDL. The following points are especially relevant for audiovisual archive content:

- Object models currently only support metadata on the level of complete objects. Intra-object descriptions, which are crucial for audiovisual content, are within the long term goals.
- Metadata formats need to include rights information and the file format and version as technical metadata. A higher level interoperability profile should not be created, but application profiles shall be harmonised using Semantic Web technologies. It is proposed that the choice of file formats follows the suggestions of the Minerva[122] and Planets projects. The use of packaging formats such as METS, MPEG-21 DIDL, XFDU is proposed for complex objects.
- Basic semantic interoperability to allow access by semantic query methods shall be established.
- Legal issues and access protection are a long term issue, but are relevant for audiovisual content.

Preservation metadata: the PREMIS data model and data dictionary define multiple levels for modelling information and structure. These are:

- Intellectual Entity: a coherent set of content that is reasonably described as a unit, for example, a particular book, map, photograph, or database
- Representation: the set of files, including structural metadata, needed for a complete and reasonable rendition of an Intellectual Entity
- File: a named and ordered sequence of bytes that is known by an operating system
- Bitstream: contiguous or non-contiguous data within a file that has meaningful common properties for preservation purposes

PrestoPRIME will extend the ability of PREMIS metadata to capture the full information needed for the 'rendition' of audiovisual signals and support all the information needed to support decode, encode and transcode with the best possible preservation of the quality (frequency response and dynamic range) of the audiovisual signals. This aspect of PREMIS will need significant extension to support audiovisual material, as it currently by definition does not encapsulate format-specific technical metadata.

The PREMIS approach currently lacks the notion of representation of the *structure* of a bitstream or of a file. Audiovisual data is not just an ordered sequence of bytes, but has structure. The mapping of data elements and structure will need to be researched, taking into account the variance and complexity of information inherent to audiovisual material. The data model will also have to support efficient handling of large and complex files and bitstreams.

PrestoPRIME will define what further information about structure is needed to mitigate against loss and to ensure that the metadata and signal quality of audiovisual material is maintained. We will implement support for the PrestoSpace data model and format[123], which is based on

[119] S. Chambers, "Towards Metadata Interoperability between Archives, Audio-Visual Archives, Museums and Libraries: What can we learn from The European Library metadata interoperability model?", EDL project D1.1, ECP-2005-CULT-38074-EDL, Aug. 2007.
[120] Discovering Music Archives, http://www.dismarc.org
[121] S. Gradmann, "Interoperability of Digital Libraries – Report on the EC working group on DL interoperability", Lisbon, Sept. 2007, http://bnd.bn.pt/seminario-conhecer-preservar/doc/Stefan%20Gradmann.pdf
[122] http://www.minervaeurope.org

documentation models at major European broadcast archives and their business processes[124]. This describes the relations between editorial objects and their realisations as materials, the relations between materials and sources, and between materials, e.g. part-of or derivations such as transcoding. The data model, expressed in XML, supports various types of annotations of an *editorial object,* including information on (a) language and identification; (b) publication and production; (c) realising material instances; (d) editorial partitioning; (e) detailed content descriptions; (f) external metadata; (g) ancillary information.

3.4.3 Data Management

The most common approach to preservation of content in audiovisual archives is to use dedicated in-house systems (ranging from tapes on shelves through to automated mass storage systems) and a programme of migration[125], i.e. moving content from one technology to another in order to address format obsolescence or the obsolescence of the hardware/software used to store or play the physical media on which the content resides.

Migration isn't the only approach[126]; preservation using emulation[127] or multivalent[128] techniques built on the UVC concept[129] are used in other domains, e.g. for scientific data as demonstrated by the FP6 IP Planets project. Planets uses a modular emulator to allow obsolete software applications to run in a simulated computer environment, and a Universal Virtual Computer, which provides an alternative approach to emulation designed to allow interaction with software long into the future. Outsourced, distributed and federated storage and content processing infrastructures offer an alternative to in-house systems[130]. The SHAMAN project[131] combines UVC and federated environments to create a next generation digital preservation environment with corresponding preservation tools for analysing, ingesting, managing, accessing and reusing information objects and data across libraries and archives. Three prototypical applications will support trials in scientific publishing, parliamentary archives, industrial design and scientific domains. SHAMAN uses a Multivalent[132] preservation architecture which preserves the ability to manipulate the encoding format of a digital entity. For a given data type, a media adaptor is built for the Multivalent browser. The Multivalent technology and media adaptor are archived. The digital entity remains unchanged, while making it possible to apply new operations that become available in new versions of the Multivalent preservation architecture.

At the same time, the digital library community has been busy creating software frameworks for implementing preservation environments. These include open source solutions, e.g. DSpace[133] which provides standard services for ingestion and access and is ported to run on top of SRB for

[123] C. Bauer, F. Rosensprung, S. Lajtos, L. Boch, P. Poncin, C. Herben-Leffring, “Analysis of current audiovisual documentation models, Mapping of current standards”, PrestoSpace Deliverable 15.1, Mar. 2005. http://www.prestospace.org/project/deliverables/D15-1_Analysis_AV_documentation_models.pdf
[124] G. Dimino, L. Boch, A. Messina, W. Bailer, C. Bauer, V. Tablan, “PrestoSpace Documentation Platform”, PrestoSpace Deliverable 15.2, v1.01, Feb. 2008.
[125] For example, see the PrestoSpace preservation wiki http://wiki.prestospace.org/
[126] For example, See the curation manual issued by the UK Digital Curation Centre for a review of preservation strategies. http://www.dcc.ac.uk/resource/curation-manual/chapters/
[127] S. Granger, "Emulation as a Digital Preservation Strategy", D-Lib Magazine 6 (10), 2000 http://www.dlib.org/dlib/october00/granger/10granger.html
[128] A No-Compromises Architecture for Digital Document Preservation Thomas A. Phelps and P.B. Watry. Proceedings of the 9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2005), September 18-23, 2005 Vienna, Austria. http://multivalent.sourceforge.net/Research/Live.pdf
[129] Long Term Preservation of Digital Information, Raymond A. Lorie, IBM Almaden Research Center. http://portal.acm.org/citation.cfm?id=379726
[130] Building Preservation Environments with Data Grid Technologies R. Moore in the American Archivist Journal, pp. 139-158, Vol. 69, No. 1, Spring/Summer 2006. http://archives.gov/era/pdf/2006-saa-moore.pdf
[131] ftp://ftp.cordis.europa.eu/pub/ist/docs/digicult/shaman_en.pdf
[132] Thomas A. Phelps and P.B. Watry, “A No-Compromises Architecture for Digital Document Preservation”, Proceedings of the 9th European Conference on Research and Advanced Technology for Digital Libraries (ECDL 2005), September 18-23, 2005 Vienna, Austria. http://multivalent.sourceforge.net/Research/Live.pdf
[133] DSpace http://www.dspace.org/

managing distributed data, Fedora[134] which associates display functions with each data type, allows relationships to be imposed on records, and maps semantic labels on records to an ontology, as well as simple, off the shelf systems such as Greenstone[135], and commercial systems including the ExLibris[136] solution used in this project.

Choice of a preservation strategy and the use of digital library tools are not sufficient in themselves to create a complete preservation environment, with a range of further tools and services[137] required to support preservation processes, e.g. format identification, selection of a preservation format, migration through transcoding, quality assessment etc. As just one example, consider JPEG2000, which was developed as an international standard for the compression of still digital images, but has evolved to support lossless and visually lossless image compression that supports multi-resolution imaging and scalable image quality, with high dynamic range and metadata support. In 2005, the Digital Cinema Initiative, a joint venture of six Hollywood movie studios, adopted JPEG 2000 as the basis for the Digital Cinema Package (DCP), which is used to distribute digital movies to theatres in an MXF wrapper containing the images, audio and other movie data. There is interest in the archive community in the use of JPEG 2000, which has been endorsed by the Digital Preservation Coalition[138] as an archival format to store “visually lossless” files[139], which can reduce the file size by an order of magnitude in comparison to uncompressed TIFF. However, it remains to be seen whether JPEG2000 is a viable preservation format for video.

3.5 Moving Digital Content into Files

In the general world of archives and digital preservation, material is either non-digital and on shelves (like books) or it is digital, manifested as files on some kind of storage (like scans of book pages stored as TIFF or JPEG files).

The audiovisual world has something else to worry about: digital content that isn’t in files:

- audio on audio CD, minidisc or R-DAT tape
- various specialist digital audio formats from the 1980’s which stored audio on videotape
- video on DVD or on digital videotape, such as Digibeta or the various DV formats[140] (the early SMPTE ‘D-Formats’: D1, D2, D3, D5 and D6 are now all obsolete – D4 never existed)

There is a general principle in archiving: keep the original (the artefact), or at least keep the bits intact (as they were on the original artefact). How does that apply to the particular problem of saving non-file-based audio and video?

There are complications, and PrestoPRIME is working (with partners) on a guideline document, being prepared for the May 2010 Joint Technical Symposium[141]. The following is a summary of that guidance:

In an attempt to uphold archive principles and ‘save the bits’, three cases occur:

[134] Fedora http://www.fedora.info/
[135] Greenstone http://www.greenstone.org/
[136] ExLibris digital library solution: http://www.exlibrisgroup.com/
[137] For example, Miguel Ferreira et al identified the following necessary tools to support migration based preservation in their automatic digital preservation system: A format identification service that also checks the integrity of digital objects; A service that produces recommendations of optimal migration options (selection of a migration option); A service to carry out format migrations (the conversion); A service to determine the amount of data loss resulting from a migration (evaluation of results); A service that provides information about the formats that are at risk of becoming obsolete.
http://www.ariadne.ac.uk/issue48/ferreira-et-al/
[138] Digital Preservation Coalition http://www.dpconline.org/graphics/index.html
[139] See the DPC technology watch report on JPEG2000 http://www.dpconline.org/docs/reports/dpctw08-01.pdf
[140] http://www.digitalpreservation.gov/formats/fdd/fdd000183.shtml
[141] http://www.jts2010.org/

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.

Case 1: the bits are not available (to the external world); minidisc, Digibeta: for both these formats (both by SONY), the bits are internally uncompressed by electronics within the playback device, and than a standard, uncompressed signal is presented to the outside world.

For minidisc, there is (or was; most minidisc equipment is out of production) at least one professional deck[142] that would clone from minidisc to minidisc without a decode, but it is unclear whether the compressed (native) bits could be captured and saved to a computer. Certainly some kind of special apparatus would be needed to 'intercept' the bits with a computer.

For Digibeta, there is no such "non-decoded data" cloning option, though even professionals in the industry are frequently unaware of the fact that the bits coming out of a Digibeta machine are NOT the bits on the digital videotape. The internet abounds with references to 'Digibeta cloning', which is an oxymoron – and even to lengthy descriptions of why to use Digibeta copies because they are an 'exact digital clone'[143].

For a massive project, an archive could attempt to modify equipment to expose the un-decoded bits, and thus allow an archive to 'keep the original'? But what then? There is no software for decoding minidisc or Digibeta encodings, because there has never been a way to get that kind of data into a file, and into a computer (except in Sony's own labs).

Late news on minidisc: there is a way to get to the bits, and there is software to play the bits, so perhaps best advice would be to now consider minidisc as Case 2, considered next.

The ffmpeg project announced on their website[144] on September 23, 2009:

> "In 1992 Sony introduced the first Minidisc player. 17 years later it is now possible to transfer and play back the raw ATRAC data from the actual digital disc with the help of FFmpeg, tools developed by the Linux Minidisc project[145] and official hardware (MZ-RH1)[146]. So if you have lots of digital recordings stored on Minidisc now is the time to archive it all."

Case 2: the bits are available, and a clone is made: CD, DAT, DV (and now minidisc!): this case is easy to justify, as it simply follows the main archive principle: preserve the artefact.

The complications are around what else needs to be done

- *usability of the clone*: CD and DAT have uncompressed audio, probably the most widely-supported form of an audio signal. However the minidisc has ATRAC coding, which was proprietary to Sony, and has only very recently had a decoder as a Linux tool on the ffmpeg website. That form of audio would only confuse people if used as a standard format for distribution from an archive, so in addition to the minidisc clone file, there would have to be a decoded file, saved as uncompressed audio, and one of more distribution files on formats like MP3 (unless distribution files are created on demand).

[142] http://www.minidisc.org/part_Sony_MDS-B5+B6P.html
[143] For instance: "Digibeta is the primary digital archival format and is considered the best preservation standard by the archival community. There is no generation loss of content when remastering from Digital Betacam because it is an exact digital clone of the original."
http://www.eai.org/resourceguide/collection/singlechannel/bestpractices.html
[144] http://ffmpeg.org/
[145] https://wiki.physik.fu-berlin.de/linux-minidisc/doku.php
[146] http://www.minidisc.org/part_Sony_MZ-RH1.html costing roughly £250 at Sony dealers in late 2009

DV files are very widely supported, so they could be used as the standard archive 'high quality' format – but have other long-term issues, considered next:

- *suitability of the clone for digital preservation*: how long will the ATRAC decoder be around? That's the same question that has to be asked for every format, **except for uncompressed!** An archive would always want to make an uncompressed version of an audio file as soon as the file arrived at the archive. For video, there is an 8:1 size difference between uncompressed and DV coding, which could be economically significant. Therefore as an instance of *temporary archiving*[147], just the DV could be saved until such time as DV is at risk of becoming unusable – at which time an uncompressed 'new master' could be made, and stored much more cheaply than at present.
- *workflow requiring an unencoded or standard signal*: if there is any automatic checking in a digitisation or digital archive workflow, it may not work on the encoding from the original an archive is seeking to preserve. Audio analysis software may very well not accept an ATRAC signal, and video checking may need an uncompressed (e.g. SDI) signal rather than a DV signal. In such a case, making the required signal (as a decode process) has to be added to the workflow, which then means that the quality checking (or whatever) is not operating on the *artefact*, but is running on a decoded signal which one hopes is pretty much the same as the artefact, and certainly needs to be frame-synchronous.

  Having saved a DV file, and produced an SDI signal for quality checking, which is the master? There are complexities here, which the principles of archiving don't cover. Really there are two masters, as the SDI has been checked, and if saved uncompressed then it's the best format for carrying the content into the future. The DV is 'the artefact but not the master' – a concept that makes as much sense as 'temporary archiving'. Principles are simple; reality is complex.

Case 3: the bits are available, but a clone is *not* made: this is the case for the D3 preservation project of the BBC (INGEX Archive[148]). The Panasonic D3 format stores an uncompressed 4:2:2 8-bit data composite signal. Standard SDI is component, not composite, and is 10-bit 4:2:2. Component vs. composite has been mentioned (*Section 2.1.3*), and component was described as being 'always best' – but the composite signal is the artefact.

The artefact could be saved, just as in Case 2, and then a fully-decoded version could be produced and saved as the 'new master'. The BBC has not 'saved the artefact', because we felt it would never be useful for any purpose, for the following reasons:

- the data is in an obsolete format: PAL-encoded composite data. There is no support for this data as a file-based encoding. There isn't actually a 'capture interface' for the data: only a hardware interface to get the data from the D3 playback machines into a PAL-decoder (and that hardware is now obsolete technology and very hard to maintain).
- all future use of the artefact would first require decoding from PAL to component. The BBC has hardware to do that conversion. There is software to do PAL decoding, but the BBC are using an advanced frequency-domain decoder which has no standard software equivalent. To do as good a job in the future as can be done today, a software transform decoder would have to be created, and kept operational into the indefinite future.

The thinking went as follows:

- the D3 project is difficult, and has holes in the workflow that need time and resources for custom development
- saving the artefact would require a lot more custom development:

[147] http://wiki.prestospace.org/pmwiki.php?n=Main.Roadmap the concept is being further developed in PrestoPRIME; documents in preparation
[148] http://ingex.sourceforge.net/archive/

  - a method to allow a computer to 'capture the bits';
  - a software PAL decoder for an obsolete format, developed only for a file which we would never expect to use

- there was great difficulty (shortage of parts, obsolete devices) interfacing the D3 machines to the PAL decoders; simultaneously connecting the D3 machines to a computer capture system could roughly double the hardware problems
- the frequency-domain decoder is, in principle, reversible; therefore there is a weak argument that says the 'artefact' could be computed from the files that we are making.

And the conclusion was to allocate resources to what had to be done, and not double the technical difficulties by also capturing and saving the artefact.

3.6 Projects

The projects developing digital preservation technology were reviewed in Section 3.3 and 3.4, above. Other projects should be mentioned, which are actually using the technology.

Possibly the biggest single digital archive project is at the US National Archives and Records Administration (NARA). This project is about records in general, not audiovisual content. Because of its size, much technology is being developed for, or adapted for, the NARA project – which makes it important for everyone, as the rest of us may find ourselves re-adapting (for our purposes) technology that we can better understand if we at least know about the NARA project. NARA is building ERA[149], the Electronic Records Archive for the USA. The project started in 2004, with a declared budget in the region of $100 million, but fundamental issues were being planned for several years before that, as summarised in an 83 page report[150] issued in 2003. Recent reports[151] indicated that about $60 million has been allocated so far.

There is extensive online documentation. For audiovisual collections, there are at least three reasons why the ERA project is important:

- a 'document life cycle' approach, as traditionally used by archives, is fully relevant to electronic documents and to audiovisual content;
- the technology required to really undertake comprehensive management of files is far from trivial. The 'brief' summary of their plans, just cited, is 80 pages, and the full list of documentation is just short of 20 major documents. A role of PrestoPRIME and the Competence Centre will be to distil such information into a form that is usable by small institutions;
- ERA builds preservation methodology into the overall management, as part of the original design of the system.

Also in the USA is the National Digital Information Infrastructure & Preservation Program (NDIIPP[152]), a Library of Congress (LOC) initiative. NDIIPP is very much aimed at digital preservation, and has various sub-projects of direct relevance to digital audiovisual content:

- the LOC online data "Sustainability of Digital Formats[153]" is part of the NDIIPP work

[149] http://www.archives.gov/era/
[150] Building an Electronic Records Archive at the National Archives and Records Administration: Recommendations for Initial Development (2003) http://www.nap.edu/openbook.php?record_id=10707
[151] http://historycoalition.org/2009/06/29/nara-fy-10-budget-clears-first-hurdle-in-the-house/
[152] http://www.digitalpreservation.gov/
[153] http://www.digitalpreservation.gov/formats/

- as with ERA, there is a huge amount of associated technology: over 30 software tools are listed in their "Tools and Services Inventory[154]", ranging from validation and data integrity tools, to web-harvesting technology and on to entire digital archives.
- WNET (the US public broadcaster in New York City), working with New York University, is implementing OAIS technology on broadcast files – as the Preserving Digital Public Television[155] project. This work is perhaps the single most important digital preservation project currently running, so far as broadcast archives are concerned. The role of PrestoPRIME will be to work with this project, and whatever comes after it, to ensure exposure of this pioneering work.

The Preserving Digital Public Television project has actually implemented the detail of OAIS, building 'submission information packages' out of the files and associated metadata arising from actual television production.

In Europe, EC project MEMORIES[156] has also implemented a form of OAIS, concentrating on audio files from radio and music collections. The MEMORIES project has also performed innovative work on semantic indexing, making the whole project rather complex. The actual implementation of OAIS relies upon knowledge of Autonomous eXchange Entities (AXE units) with the overall AXIS structure, which are constructs associated with the European Media Wrapper and the TITAN[157] initiative.

Finally, the EDCine project has developed a comprehensive strategy for both distribution and archiving of digital film. They use MXF as a wrapper and two versions of JPEG2000 as the codec: lossless for archiving, lossy for distribution. The lossy JPEG2000 codec has the advantage (not exclusive to JPEG2000, but certainly not available in MPEG-family codecs) that it efficiently supports producing proxies (access copies) at lower bit rates. This feature allows an archive to hold one 'mezzanine' version, and generate copies at any lower quality and bit-rate, efficiently and on demand.

3.7 Role of PrestoPRIME

Audiovisual issues that are outside standard digital preservation technology include:

- a time dimension: all the data within audiovisual files has a time dimension, and much of the metadata may also be time-dependent (e.g. time code, subtitles, rights and provenance information about segments).
- a signal representation dimension: the numbers come from, and will be turned back into, one or more signals, whose essential quality parameters are frequency range and dynamic range. The evolution of the data, through migration, emulation and transcoding (or any combination) needs to be designed from a signal-processing perspective (preserving the signal), not just from a data perspective (preserving the numbers), in order to preserve as much as possible of the original signal quality.
- a multiplicity of encodings: all signals are 'encoded' in order to be represented by digits, but audiovisual encodings divide into uncompressed and compressed, lossy compression being an absolute requirement for practical distribution and access. All digital broadcasting uses compression (roughly 5:1 for audio and 50:1 for video); web access typically uses another order of magnitude of compression. These various 'proxies' form a family of files, whose relative provenance, rights, quality and evolution over time need 'collective management'. This situation differs in kind from the standard model for digital libraries,

[154] http://www.digitalpreservation.gov/partners/resources/tools/index.html
[155] http://www.thirteen.org/ptvdigitalarchive/
[156] http://www.memories-project.eu/index.html
[157] http://www.titan.be/

based on a single 'original', with an inherent assumption that viewing copies can be made on demand. Two obvious differences are:

- No 'original': much video data is archived in a compressed form. Compression algorithms have a lifespan in decades, at best. MPEG-2, the basis of digital broadcasting and DVDs, is already being superseded by MPEG-4. The problem with 'no original' is that it creates an open question: how best to get from one encoded form to another. The obvious route is to decode to uncompressed and start again, but this is not necessarily the best answer. Decoding produces an "uncompressed proxy", but it does *not* produce an 'original'. Work on MPEG-2 by the BBC[158] (and partners, in EC-supported project ATLANTIC) has shown that lower loss of signal quality can be achieved by staying within the compressed domain. What has not been shown is how to move – optimally – from one compressed format to another.
- Migration of viewing copies: even if a higher-quality or uncompressed version is available, the cost of re-encoding from the original may be significantly higher than the cost of an efficient form of transcoding (moving directly from one encoding to another, without moving back to an uncompressed representation).

Some of the issues just listed affect access, and will be covered in the next section. Regarding digital preservation technology, PrestoPRIME has produced an overall strategy, including looking at the roles of migration, emulation and 'temporary archiving', and specifically covering audiovisual content. This deliverable (D2.1.1 Preservation Strategies) will be available in early 2010.

[158] http://www.bbc.co.uk/atlantic/

4 Access

The payoff of preservation is access. The last section looked at a new set of problems: the digital preservation of file-based content. This section looks at a new range of solutions provided by file-based content in digital libraries:

- technology: files don't need audio and video playback machines;
- location: file-based content can be access via the Internet;
- discovery: content *aggregated* in a well-known digital archive – like Europeana – will greatly increase the probability of success in connecting content to people who want to find that content.

4.1 Audiovisual material in digital libraries

Time-based media has special properties (a time dimension, to state the obvious one). These properties deserve support (within digital library system) that pays attention to audiovisual requirements, which fall into two categories: function and metadata:

- **function:** things that can be done with audio and video
- **metadata:** the documentation needs of audio and video

4.1.1 Function

Here are six areas where audio and video have specific ways of operating, that are different from text documents – and that may not be well supported by digital libraries:

- **segmentation and granularity** – books and other documents aren't just a continuous stream of text: they have structure, consisting of division into various units (chapters, paragraphs; body, references, appendix; pages). Video also has structure (a news programme would divide down into items, shots and individual frames), but (generally) there is no table of contents or other obvious textual description of the structure. Asset management systems often have a story-board or light-table display, to represent video by a sequence of images (key frames). There are many other experimental methods for representing the structure of video along a time dimension[159]. Because audiovisual files can be very large (DVD quality video can be several gigabytes per hour), people using web access would benefit from a way to see a structure, and navigate within that structure, without having to first download an entire huge file.

- **time-based metadata** – the information provided in text documents by a table of contents and chapter titles and other section labels, would be provided on audio and video by metadata that is 'time-stamped' or somehow attached to a time-line, attached to the temporal dimension of the audio and video. This functionality is easy to imagine, but not at all easy to find[160]!

- **time-based navigation** for retrieval/playback – given some kind of segmentation of audio and video, and some kind of time-based metadata for providing information about individual segments, and given some way to link time start-stop information to the operation of a media player – the user can at long last move around in an audiovisual file, getting quickly and precisely from one point to another.

- **citation** seems a very simple issue. A researcher who wants to quote a text, precisely, just has to give a page number – or even page, paragraph, sentence and word if greater

[159] Roberto Basili, Marco Cammisa, Laurent Boch, Alberto Messina, Giorgio Dimino, Valentin Tablan, Borislav Popov, Werner Bailer, Walter Allasia, Michele Vigilante "From Video Segmentation to Semantic Indexing: The PrestoSpace Approach" Proceedings of ESA-EUSC 2006: Image Information Mining for Security and Intelligence http://www.joanneum.at/no_cache/en/jr/publications.html?tx_publicationlibrary_pi1[showUid]=4458

[160] At this 2009 conference, time-based metadata is still seen as quite visionary: http://www.streamingmedia.com/article.asp?id=11260

precision is needed. How is audio and video cited? First the user has to have a time-code or some other pointer to content, and second that pointer has to be exportable (from the application being used by the researcher), publishable (so it can go into a report or thesis or just into an email), and usable by someone else, at some other time, using some other viewing application on the same audiovisual file. All that needs standards which are not yet in place. We can send URLs to pinpoint a file, but time-code information is either:

- publishable but not really usable: reading the minutes and seconds from the display of a player gives easily publishable numbers, but no way to use those numbers to get another application to start playing at that specific point; the new user would have to manually move to that time.
- usable within one application (like the special codes that can start a YouTube clip at a specific point) but neither publishable as general information, nor usable by any other application. An aggregator like YouTube, that holds the content, may not be bothered about compatibility issues – YouTube operates a kind of closed universe. An aggregator such as Europeana, where the catalogue "clicks through" to content on other systems, possibly using players from whatever tools are in a particular users' web browser, has a huge problem with getting a 'simple citation' to actually work.

- **time-based annotation,** including user/viewer annotation and tagging: citation is just a pointer to a place in a file. Annotation is adding some sort of text to that pointer. The possibilities and complexities then fall into various cases:
  - annotation by the 'owner' of the content, possibly added once-only when content was placed into some sort of repository, or with more functionality annotations could be added later;
  - annotation by users of the content, so definitely added afterwards;
  - controls on who can see annotations;
  - adding annotation text to the rest of the metadata associated with an item, so that annotations could support search (augmenting formal indexing, full-text search and analytical cataloguing)

- **communities:** users don't just *look* at information on the web, they *use* it – an area of activity that includes social networking tools:
  - time-based user annotation: notes on a specific part of a file, that can be shared;
  - time-based user tagging: not just saying what the whole item is about, but identifying specific time points or durations. Ideally tagging for video would be not only temporal but spatial: identifying specific parts of an image.
  - time-based recommendations (a citation, really, but compatible with bookmarking and tagging tools such as Delicious, Digg, reddit or with social networking sites such as Facebook);
  - creating a clip or list of clips (edit decision list) and passing that to someone else, or to a community of users.

4.1.2 Metadata: Interoperability for Access

The preceding subsection was labelled *function*, but made much mention of metadata. That section was discussing metadata in terms of functionality that it either needed or enabled. This section is entirely about metadata, focusing not on specific functions but on the general issue of how metadata can be made to work across the diversity of standards and sytems that constitute the Internet. The following problems are common to all content going into digital libraries, with one complication: for audiovisual, it's worse – because audiovisual content is really just entering the digital library world.

Many portals offer access to cultural digital content, targeting different user groups: the VideoActive project[161] has collected a list of over a hundred online access services to audiovisual archives and libraries[162]. The EDLnet project portal *Europeana* is intended to become the future reference point for access to cultural and audiovisual contents. But for large-scale portals to work, a high degree of interoperability between archives and collections is essential to simplify the process of including new material. WorldCat[163] and ArchiveGrid[164] are portals for libraries and archives respectively, provided by the Online Computer Library Center (OCLC)[165] linking thousands of institutions using MAchine-Readable Cataloging (MARC)[166] records as the input format. The MICHAEL (Multilingual Inventory of Cultural Heritage in Europe) project is facilitating access to cultural heritage information by producing an inventory of digital collections.[167]

Exchanging metadata is the key to ensuring access to audiovisual collections, establishing interoperability among audiovisual collections, and between audiovisual collection and other cultural heritage institutions. Metadata exchange is hindered by the diversity of metadata formats and standards that exist in the media production process and in different communities. Metadata interoperability needs to be established between different parties involved :

- *professional content providers and users (e.g. broadcasters, archives, libraries, production houses)*
- *professional content providers and consumers (e.g. adapting metadata delivered on the Web, to mobile devices)*
- *users contributing content and metadata and professional content providers (e.g. metadata of user generated content, user generated annotations, relational information provided by users, such as Wikipedia articles)*.

and on two levels:

- *syntactic* interoperability: metadata can be accessed and processed in the same syntactic format, typically some XML format. Note that this does not imply that all metadata are XML data, only that they can be rendered as such (with services or wrappers). RDF[168] is the Web standard with an XML syntax designed for achieving syntactic metadata interoperability.
- *semantic* interoperability: metadata can (partially) be interpreted within the same semantic frame of reference. Meaning of metadata of one archive (typically coded in in-house metadata vocabularies) needs to be linked with metadata from another archive. Thus, it requires alignment of archive vocabularies, which are partial as vocabularies differ in scope and perspective.

[161] http://videoactive.wordpress.com
[162] VideoActive bookmark collection of audiovisual archives offering online access, http://del.icio.us/VideoActive
[163] http://www.oclc.org/worldcat/
[164] http://www.archivegrid.org/web/index.jsp
[165] http://www.oclc.org
[166] http://www.loc.gov/marc/
[167] http://www.michael-culture.eu/
[168] Resource Description Framework (RDF). http://www.w3.org/RDF/

The Open Archives Initiative[169] develops and promotes technologies for archive interoperability: the *Protocol for Metadata Harvesting* (OAI-PMH) and *Object Reuse and Exchange* (OIA-ORE). OAI-PMH is a mechanism for repository interoperability that can be used to exchange documents according to any XML format as long as it is defined by XML schema. The international OAI-ORE effort works towards a solution based on publishing Resource Maps that describe compound objects, referencing resources in their compound object context, and mechanisms to facilitate discovery of Resource Maps[170]. Search and Retrieve by URL (SRU)[171] is a protocol for XML-focused Internet search, which is among the protocols used for Europeana[172]. SRW is a variant that uses Web services instead of URLs for transporting the query.

We should pause here, and reflect that all the technology described in the last several paragraphs was developed for text, not for time-based media. Harvesting, aggegation and search/retrieval all have well-developed protocols, but use of these protocols on time-based metadata is in many cases impossible, and in other cases only very recently developed. The author knows of only one time-based annotation system that has succeeded in exported time-based metadata using OAI-PMH – the Lignes de Temps system from IRCAM in Paris[173].

There are two aspects of metadata interoperability with the Semantic Web:

- providing an interface for Semantic Web agents to access the content portals
- using Semantic Web technologies.

Both in the EDL project report on metadata interoperability[174] and in the Bricks project[175] the use of Semantic Web technologies is proposed as a way of mapping between metadata schemes without defining specific converters or a “super-scheme”. The EC working group on digital library interoperability[176] defines Semantic Web interoperability with the outside world as one of the goals. In the MultiMatch project[177], OWL (Web Ontology Language[178]) is used as a representation of the internal metadata model, which can also serve as a gateway to the Semantic Web.

The representation of multimedia metadata in formats that are interoperable with the Semantic Web is still an active research issue. A number of multimedia ontologies have been proposed, partly defining new metadata schemes, partly representing existing ones (e.g. MPEG-7). A good overview on the work on multimedia ontologies has been produced by aceMedia[179]. COMM[180] and DOME[181] are recent proposals for new multimedia ontologies.

[169] http://www.openarchives.org/
[170] H. Van de Sompel, C. Lagoze, “Interoperability for the Discovery, Use, and Re-Use of Units of Scholarly Communication”, CTWatch Quarterly, vol 3, nr. 3, http://tinyurl.com/27grpo/
[171] http://www.loc.gov/standards/sru/
[172] S. Gradmann, “Digital Library Interoperability technical and object modelling aspects of Europeana”, Frankfurt, 2008, http://www.edlproject.eu/conference/downloads/EDLconf_Gradmann.pdf/
[173] Lignes de Temps http://www.iri.centrepompidou.fr/ (and scroll down to Lignes de Temps)
[174] S. Chambers, “Towards Metadata Interoperability between Archives, Audio-Visual Archives, Museums and Libraries: What can we learn from The European Library metadata interoperability model?”, EDL project D1.1, ECP-2005-CULT-38074-EDL, Aug. 2007.
[175] http://www.brickscommunity.org/
[176] S. Gradmann, “Interoperability of Digital Libraries – Report on the EC working group on DL interoperability”, Lisbon, Sept. 2007, http://bnd.bn.pt/seminario-conhecer-preservar/doc/Stefan%20Gradmann.pdf
[177] Multilingual/Multimedia Access To Cultural Heritage, http://www.multimatch.org/
[178] OWL: The character Owl in the Christopher Robin books (A A Milne) had his name misspelled as WOL, so the Web Ontology Language http://www.w3.org/TR/owl-features/ people did the same, in reverse. To quote Dorothy Parker: “At this point, Tonstant Weader fwowed up” http://www.merriam-webster.com/cgi-bin/wftwarch.pl?011806.
[179] H. Eleftherohorinou, V. Zervaki, A. Gounaris, V. Papastathis, Y. Kompatsiaris and P. Hobson, “Towards a Common Multimedia Ontology Framework (Analysis of the Contributions to Call for a Common multimedia Ontology Framework Requirements)”, AceMedia Technical Report, Apr. 2006, http://www.acemedia.org/aceMedia/reference/multimedia_ontology/index.html
[180] http://comm.semanticweb.org/
[181] http://www.micc.unifi.it/dome/

The sheer amount of metadata resulting from the fine-grained description of multimedia content can be a limiting factor. In a scenario where just the *visual* modality is described by low-level descriptors of key frames, one million triples are required to represent a single hour of video[182]. Semantic Web technologies cannot easily be applied to all the metadata in a repository: very careful consideration is needed to decide which metadata should be represented in a Semantic Web compatible format. An approach for deploying multimedia metadata on the Semantic Web (using any existing ontology representation for the multimedia metadata) has been proposed by Joanneum Research[183], with an application to the cultural heritage domain[184].

Thesauri are useful for indexing and retrieval on the Semantic Web, but they are often not published in RDF/OWL. Moreover, different organisations use different thesauri (cf. the PrestoSpace report on documentation models of European archives[185]). A structured method is required to convert thesauri to RDF for use in Semantic Web applications and to ensure the quality and utility of the conversion. Moreover, if different thesauri are to be interoperable without complicated mappings, a standard schema is required. The Web standard SKOS[186] is attractive because it offers syntactic interoperability (through RDF) as well as a limited form of semantic interoperability (through its predefined semantic vocabulary relations).

4.2 State-of-the-Art

A great deal of state-of-the art has already been mentioned in Section 4.1. This section will not look at all the applications which collect or otherwise provide access to audiovisual material, but instead will just look at the use of technology that supports finding audiovisual content.

The different uses of metadata to support search can be put into five categories:

- **Informal user-indexing**: the major example is YouTube, which has no formal indexing, but it does support tagging (by the uploader, not by viewers). Almost everything exists on YouTube, but there is no guarantee that anything specific can be found. One reason for the development of *social-networking* tools must have been sheer frustration at application that had no internal tools. People don't so much find content on YouTube, as have it pointed out to them through recommendation/bookmark systems, or through very specific *external* (user-generated) tagging using systems such as Delicious[187]. In mid-2009 YouTube added annotation, again only as something that can be added by the person supplying the video, not by general viewers. Viewers can leave comments, but those apply to the whole file, not specific points (though of course there is nothing to stop people type "minutes and seconds" information into the comment).

  There are mechanisms for supplying start-stop times with a YouTube URL, so that a referenced item will start playing at a specified point within a file, and for a specified duration. However these methods are in no way a standard, and only applies to files on

[182] http://lists.w3.org/Archives/Public/public-xg-mmsem/2007Jan/0001.html
[183] RDFa-deployed Multimedia Metadata (ramm.x), Specification, 2007. URL: http://sw.joanneum.at/rammx
[184] M. Hausenblas, W. Bailer and H. Mayer, "Deploying Multimedia Metadata in Cultural Heritage on the Semantic Web," in First International Workshop on Cultural Heritage on the Semantic Web, co-located with the 6th International Semantic Web Conference (ISWC07), Busan, KR, Nov. 2007.
[185] C. Bauer, F. Rosensprung, S. Lajtos, L. Boch, P. Poncin, C. Herben-Leffring, "Analysis of current audiovisual documentation models, Mapping of current standards", PrestoSpace Deliverable 15.1, Mar. 2005. URL: http://www.prestospace.org/project/deliverables/D15-1_Analysis_AV_documentation_models.pdf
[186] Simple Knowledge Organization System (SKOS). http://www.w3.org/2004/02/skos/
[187] http://delicious.com/

YouTube. They rely on Flash-player functionality, one using a 'start' command[188], and the other uses 'watch'[189].

- **Evaluated user-indexing**: there are now various "tagging games" that reward the quality of user-generated indexing. The basic idea is consensus: if terms from two people agree, that gets rewarded. The time-based information associated with the tags comes from requiring the tagging to be done in real time while watching the video, so that the tags apply to some duration just before the typing of the tag. Yahoo developed one such system[190] for the Semedia[191] project, and the Dutch Audiovisual Archive uses another system, WAISDA[192]. Semedia actually developed an impressive range of tools for audiovisual indexing and access, showcased here: http://www.semedia.org/showcase.html

- **Automatic indexing**: YouTube is the major aggregator (by a huge margin) for user-generated content (or at least, user-uploaded content). However major broadcasters produce (collectively) thousands of hours per day of radio and television material that also goes 'on the web', and aggregators have developed who provide ways to find material from all that content. They do it through text: text on the web pages where the content was found, and text included with video as subtitling (closed captioning). Blinx[193] is a leading example of the approach, but by no means the only one: a search for "video search engine" found 9 million hits. Blinx points to 35 million hours of content, but Fooooo (which the author has only just discovered) claims 300 million hours, mainly from YouTube. All the major search engines (Google, Yahoo, AltaVista ...) now offer video search functionality.

  PrestoSpace demonstrated a form of automatic indexing that didn't rely on subtitles, but could use speech recognition to find basic terms, which were then used as query terms in a web search. For recent news material, online newspaper text (or broadcast news websites) could then be searched to get large amounts of text, which could then be process for 'entity extraction' by Natural Language Processing tools. All this was wrapped up by mapping the discovered entities to an ontology, giving some of the functionality of traditional hierarchical classification – but totally automatically (Ref 159).

- **No indexing needed**: for over 15 years there have been systems which sought to find content without any associated text, using *content-based retrieval*[194]. An image could be used to request the system to "find some more images like this one" – or the system could be categorised in terms that were measured on the content, though described in words: "get me something with red in the corner" – or the system could, with luck, find objects, allowing a search such as "find a sunset". All these approaches have generally (in the author's view) fallen into the category of 'solutions looking for a problem' – though there has been sufficient interest to motivate the TRECVID[195] research and evaluation, and there may be increasing interest in the 'find me more like this' approach, now that there is so much content online to form a starting point.

- **Formal manual indexing**: VideoActive (ref 161) is a project putting television archive content online, for (primarily) research by historians. This was a serious project, and the

[188] YouTube playback can use a 'start' command to start n seconds into a file: http://www.jakeludington.com/youtube/20090305_start_youtube_video_automatically_from_a_specific_time.html

[189] Add the minute and second where you want the video to begin playing to the end of the web address (URL). The format is: #t=2m27s. This URL begins at "2 minutes 27 seconds in: http://www.youtube.com/watch?v=WmxT21uFRwM#t=2m27s

[190] http://research.yahoo.com/pub/2157

[191] http://www.semedia.org/

[192] http://research.imagesforthefuture.org/index.php/tag/social-tagging/

[193] http://www.blinkx.com/

[194] W. Niblack, R. Barber, W. Equitz, M. Flickner, E. Glasman, D. Petkovic, P. Yanker, C. Faloutsos, G. Taubin, ``The QBIC Project: Querying Images by Content Using Color Texture, and Shape," Proc. SPIE, vol. 1908, Storage Retrieval for Image and Video Databases, 1993, pp. 173-187 http://www.research.ibm.com/networked_data_systems/spire/research.html#Niblack93

[195] http://www-nlpir.nist.gov/projects/trecvid/

solution to making content findable was to use traditional metadata – coming from the catalogues of the institutions which provided the content. There was a mapping to get all the metadata into a common schema (Dublin Core[196]), and the mapping was done in various ways, ranging from fully-automatic to fully manual. Obviously the manual approach will not scale to millions of items (the size of Europeana, for instance).

There are general technologies for mapping multiple classification systems onto one common system, and PrestoPRIME (University of Amsterdam, Joanneum Research, University of Liverpool) has expertise in this area. One of the outcomes of the semantic web part of PrestoPRIME (in workpackage four) will be tools for mapping multiple forms of legacy metadata into a unified ontology, represented by RDF (ref 168) technology.

4.3 Europeana, VideoActive and EUscreen

The European Digital Library (EDL) is the main partner (of PrestoPRIME and the Networked Audiovisual Competence Centre) in promoting access to content. EDL will provide guidelines on how to make AV content accessible via their Europeana portal as well as describe the role of other European 'aggregators'. Two other EU Projects, EUscreen and European Film Gateway are also providing portals both of which will be actively promoted on the site as well as any other aggregation efforts.[197]

Title	Description
D6.2.2 "European Digital Library Implementation Guidelines for Audiovisual Archives"	This report will include precise and practical guidelines to make content available via Europeana as well as an overview of developments in EUscreen and EFG.
Information on Audiovisual Portals and Aggregation Sites	This area will provide updates on the development of other audiovisual portals and aggregation efforts as well as links to the portals themselves.

4.4 Role of PrestoPRIME

In the area of access, PrestoPRIME will:

- develop tools for access to audiovisual files
- implement those tools within Europeana, or at least 'reachable' by Europeana
- test the tools using Europeana
- document the processes required to prepare media, submit it, use tools on it, and gather the results of that use (e.g. user metadata)
- provide the tools and information using the Competence Centre

PrestoPRIME will advance the state of the art with a novel approach to metadata mapping, formalising the semantics of the metadata standards involved. The formalisation will relate the common concepts to their respective manifestations in the different standards to derive mappings for a given pair of standards. The complexity of the problem is thus reduced to defining a formalisation for each standard, as opposed to a mapping between each pair of standards

Most work on interoperability and access to digital heritage collections focuses on objects that are documented on a per object basis, with little support for intra-object description. PrestoPRIME will develop approaches to cover the richness of audiovisual content, with its temporal dimension, complex internal structures and multiple versions.

PrestoPRIME will use the results on cross-language retrieval achieved in PrestoSpace and other initiatives such as TrebleCLEF[198]. As PrestoPRIME does not build its own retrieval system, there

[196] http://dublincore.org/
[197] Projects such as EUscreen, www.euscreen.eu and European Film Gateway, www.europeanfilmgateway.eu
[198] http://www.trebleclef.eu/

will not be a component for multilingual retrieval, but the metadata-related tasks in PrestoPRIME will take the requirements of multilingualism into account.

Finally, as just mention in *Section 4.2 State-of-the-Art, o*ne of the outcomes of the semantic web part of PrestoPRIME (in workpackage four) will be tools for mapping multiple forms of legacy metadata into a unified ontology, represented by RDF (ref 168) technology.

5 Conclusions

The major conclusion is that there is significant digital library and digital preservation technology for file-based content, but:

1) **specific tools usually don't work on professional audiovisual files;**
2) **there is very little use of the general technology within broadcasting**, though the situation is better in national audiovisual collections.
3) **web technology solves the technical issues** that have limited access to audiovisual archives**,** and **digitisation solves the logistical issues;**
4) **formal online access, through *digital libraries,* does not have the tools to support time-based content;**
5) **rights issues remain the major unsolved problem** limiting public access to the archives of public service broadcasters and national audiovisual collections.

5.1 Work in PrestoPRIME will:

- build on a long history of technology relating to film and video restoration to develop **file checking and quality control tools** that could make a real difference to the current cost and time bottleneck associated with manual checking (p 14)
- develop, implement and deploy **time-based tools for audiovisual access** (p19)
- develop and extend **preservation metadata**, with particular attention to PREMIS (p34)
- define the further **information about structure needed to mitigate against loss** (p35)
- **work with major projects** such as Preserving Digital Public Television **on dissemination** of results, to increase the impact and benefit of these projects (pp39-40)
- for **access**, PrestoPRIME will (p48):
    - develop tools for access to audiovisual files
    - implement those tools within Europeana, or at least 'reachable' by Europeana
    - test the tools using Europeana
    - document the processes required to prepare media, submit it, use tools on it, and gather the results of that use (e.g. user metadata)
    - provide the tools and information using the Competence Centre
- use the results on **cross-language retrieval** achieved in PrestoSpace and other initiatives such as TrebleCLEF (p49)

5.2 Tasks of the Audiovisual Competence Centre

- Providing 'technical watch reports', one of which could be about the **R-DAT equipment problems**, so people would know how to minimise the effects of the problem (p10)
- Various issues arise that really call for action at a European and global level, and the **equipment and skills shortage** (in analogue videotape playback) is one (p13)
- Helping to keep companies that provide **specialist preservation services** in business, by connecting them to users (p13)
- There is very little **training**, and that is scattered and transient (courses and initiatives come and go). Again, a major role could be played by an *Audiovisual Competence Centre*, to collect such information, keep it up-to-date and put it where people can find it. Above all, there is a need to coordinate all the available resources (from professional associations like IASA, FIAT and FOCAL, national training bodies like Skillset, universities and other formal training institutions, individual major institutions such as the British Library, and the various national film archives) to create well-planned, low-cost and above all *frequent* training courses, around Europe and around the world (p17)

- **Funding:** to gather details of the *Dutch business case* supporting Images for the Future, and make sure that information is widely and easily available (p18)
- **All of Section 2.4.2. Competence Centre Support for Digitisation** (p20)
- White papers on **major digital preservation projects** worldwide: the Competence Centre will distil such information into a form that is usable by small institutions (p39)
- **Partnership with Europeana**, and guidance on how archives can work with major content portals and aggregators (p48)
- Provide **information on PrestoPRIME tools**, and other technology (p48)

Glossary

Term	Definition
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)
content-based retrieval	Using images to find images, and its parallels in other media
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
preservation metadata	Metadata specifically about the preservation needs of a file, e.g. PREMIS
significant properties	The aspects or dimensions or qualities of digital content that need to be preserved
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.

Annex I – DPE briefing paper

Reprinted with permission from:
http://www.digitalpreservationeurope.eu/publications/briefs/ :
http://www.digitalpreservationeurope.eu/publications/briefs/audiovisual_v3.pdf

digital preservation europe

Preservation of Digital Audiovisual Content

The audiovisual (AV) record of the 20th century is at risk, and digitisation has been a solution, which has created a new problem: preservation of digital AV content. These files have requirements (size; specific formats) not adequately addressed by current technology. Best practice can be recommended, but three major changes are needed: 1) AV collections should use existing digital library and digital preservation technology; 2) that technology should advance, to support time-based media; 3) mass storage and general information technology should advance, to support specific requirements of AV files.

Problems of digital audiovisual preservation

Europe's AV holdings (in archives or other formal collections) have been estimated at 50 million hours of audio, video and film, most of it on analogue formats. About 70% of this material is at risk, and all of it will be at risk within 30 years – owing to obsolescence, deterioration and obsolete formats.

Major programmes of digitisation have started: an estimated 10 million hours has been digitised in the last decade. While AV collections have been busy changing their tapes and gramophone records into files – as a preservation solution – the rest of the world has become aware that digital files present their own preservation problems.

Large collections of files are a technical management problem; the solution is *digital library technology*. Files need maintenance: they must be named, moved to new storage (frequently!), copied for access, encoded for changing access needs, checked for validity. They need metadata actions, ranging from cataloguing to automated harvesting (for standardised and global access). Manual maintenance is simply impossible – and too error-prone – once collections reach a certain size. Digital library technology supplies automation tools for creation, maintenance and access requirements of large collections of files. There are many guides to digital library technology.

Two worlds: digital library technology comes from the academic library world. AV collections are largely outside that world. The biggest holders of content are broadcasters, and other major holdings are in film museums and other cultural and heritage institutions (one of the biggest film collections in the UK is at the Imperial War Museum). Broadcasters vary, but it is common for the computer and technical staff of a broadcaster, and the management who decide and fund technology issues – to know absolutely nothing of academic libraries and digital library technology.

The first hurdle faced in preserving AV files is to know about, understand, fund and use the existing digital library tools that change a heap of files into a managed collection.

The second hurdle is recognising that digital library tools provide management (so files can be accessed and don't get lost), but do not cover preservation. Files face a range of obsolescence issues, addressed by *digital preservation technology* – methods for ensuring that obsolete files can migrate to new standards and formats, methods for emulating old IT environments to extend the lifetime of obsolete formats, criteria for evaluating the reliability of a digital repository, and finally an overall methodology: OAIS. AV collections have difficulty finding anyone on their IT staff who has even heard of OAIS, which rather limits support for funding and implementation. Fortunately, EC project MEMORIES is developing OAIS and related procedures specifically for audio and video collections.

The third hurdle is that the specific needs of AV files are not fully supported by digital library and digital preservation technology, as discussed next.

briefing paper

digital preservation europe

References and Further Information

Audiovisual Status Surveys:
PRESTO:
http://presto.joanneum.ac.at
PrestoSPACE:
http://www.prestospace.eu
TAPE:
http://www.tape-online.net/survey.html
http://www.tapeonline.net/docs/audiovisual_research_collections.pdf
General Guides to Preservation:
http://www.bbcarchive.org.uk/
http://digitalpreservation.ssl.co.uk/

A general list of digital curation tools:
http://twiki.dcc.rl.ac.uk/bin/view/Main/DevelopmentToolList
JHOVE:
http://hul.harvard.edu/jhove/
DROID:
http://droid.sourceforge.net/wiki/index.php/Introduction
PRONOM:
http://www.nationalarchives.gov.uk/pronom
National Library of New Zealand Library metadata extractor:
http://meta-extractor.sourceforge.net/
OAI:
http://www.openarchives.org/OAI/openarchivesprotocol.html
OAIS:
http://www.dpconline.org/docs/lavoie_OAIS.pdf
Migration:
http://www.library.cornell.edu/iris/migration/
Emulation:
http://www.dlib.org/dlib/october00/granger/10granger.html
Repository evaluation criteria reference:
http://journals.tdl.org/jodi/article/view/199/180

Projects and initiatives:
MEMORIES:
http://www.memories-project.eu/
European Digital Library:
http://www.europeana.eu/

Formats:
MXF:
http://www.digitalpreservation.gov/formats/fdd/fdd000013.shtml
WAV format specification:
http://www.digitalpreservation.gov/formats/fdd/fdd000001.shtml

Presentations:
Video Formatting and Preservation, Carl Fleischhauer, NDIIPP DLF Forum, Philadelphia, 6 November 2007
http://www.diglib.org/forums/fall2007/presentations/Fleischhauer.pdf

Unique problems of digital audiovisual data and files

Because of the two worlds problem, professional broadcast formats (MXF in particular) are unsupported by many digital library and preservation tools. Other 'standard' formats are better supported, but many (eg AVI, WMV) are proprietary, which is in itself a preservation problem.

The remaining problems relate to the actual content of the files.

- most AV files are compressed. Whatever 'original quality' was lost in compression, will remain lost. Preservation should maximise retention of quality, a capability that needs to be defined and added to current technology.
- Time-based content needs tools with a time dimension (cataloguing, navigation, edit)
- The files are complex. Indeed the concept of a *wrapper* was developed to recognise the complexity of a typical AV file: multiple signals, multiple kinds of metadata – including time-domain (subtitles) and numerical (time code)
- AV preservation involves many related files: lossless and lossy encodings, multiple proxies (supporting access in multiple formats eg Real, Windows Media, MPEG, AVI, Quicktime, Flash), various stages of edit and recombination, and a range of rights information: multiple interested parties, multiple collection agencies, non-uniformity from country to country. A complex of information representing signal, metadata and rights must be preserved.

Access

Libraries have a tradition of unified access: union catalogues based on standardised metadata, to provide an 'any book, anywhere' service. Many audiovisual collections have a tradition of being closed, or open only for professional or commercial access.

Digital libraries continue the tradition of expanded and unified access, often on a national or multi-national scale, as with the European Digital Library. AV collections need the technology of digital libraries, to be accessible through major projects such as EDL. In turn, these digital libraries need to put more effort into understanding the problems of digital audiovisual data and files just discussed. In particular, digital libraries need tools for time-based accessed to both the AV signal and to the metadata (rights, for instance, can vary from moment-to-moment within a single AV file).

Much AV content is held by institutions with no history of working with libraries and who may prefer to limit access to “their” content. Marketing, branding and rights issues impede a “European Audiovisual Portal”. EDL may never include BBC content.

What to do

Despite the problems, some clear statements that can be made about AV preservation:

- **preserve the artefact:** keep the 'original', even if compressed. 'Preserve the bits', whatever else is done. AV content has one advantage: there is a lot of it, in a relatively small number of formats. Methods to 'play the bits' may exist.
- **decode to uncompressed** and save as uncompressed (*in addition* to keeping the original). This is a demanding requirement for video (100 GB/hr for 625-line TV), but storage is now very cheap.
- **enhance the metadata:** A file extension (eg .wav, .avi is not sufficient). There are over 50 registered variants of encoding within the definition of .wav; MPEG-1 and MPEG-2 use the extension ,mpg. Ideally there will be a metadata extraction tool; otherwise manual testing and documentation is needed.
- **you are not alone:** use the file-type registries, software repositories, emulation platforms, and Preservation Guides listed in the references.

Richard Wright, *Research & Innovation*, BBC
richard.wright@bbc.co.uk