Archivists of the World: You Have Nothing to Lose but Your Shelves!

Richard Wright; BBC Research & Innovation; London, UK

Abstract

Audiovisual archives have to digitise to survive, and their digital content will be enormous: 100 times larger than all other forms of digital archive content. Can it be done? This paper describes the problem, presents practical solutions -- and shows how all archives, and archive digitization projects, can benefit from the experience in the audiovisual sector.

The Audiovisual Preservation Problem

At present, European audiovisual archives hold an estimated 50 million hours of material, and the global estimate is 200 million hours [1]. **This material is valuable**: the audiovisual record of the 20th Century – the first century ever to have such documentation. A significant part is professional, such as the five million hours found in ten broadcast archives by the Presto project [2]. Material in national, university and specialist collections has specifically been selected for value, and material in commercial collections is also specifically selected for value – though with a definite definition of 'value'. Video footage sold for national or international broadcast purposes is priced at rates in the ten to 100 Euros region – per second.

The general estimate from Presto is that 70% of this material is under immediate threat: chemical decay, physical damage, obsolescence of formats and of playback equipment. All this material, except for some forms of film, will be threatened within a few decades.

Film can last for centuries – if kept frozen and not used. The 35mm format has lasted for nearly a century. It is a matter of speculation whether current film projection technology will last another century, or will be made obsolete (within two or three decades) by digital projection.

The huge amount of personal film, video and audio is not even considered in this account, though that is most at risk.

Size of the Problem, Size of the Solution

In 2006, a larger project – PrestoSpace [4] – reported on a survey covering 11 countries, with additional data from the public websites of archives in another 9 countries.

The basic questions asked were:
- How much material is in audiovisual collections?
- What condition is it in?
- What is being done about its preservation?
- What are the major problems?

How Much? In 31 institutions (out of many hundreds, but including many of the biggest) across 20 EC countries, we found 20 million individual items of film, video and audio.

What Condition Is It In? They did not know. Lack of condition assessment procedures is a main finding of the PrestoSpace User Requirements Survey.
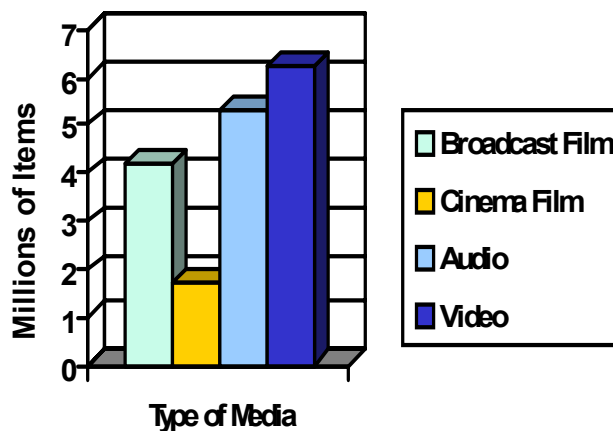

Figure 1. Audiovisual Content in Europe: PrestoSpace Survey of 30 Archives

What Is Being Done? Preservation projects were planned or underway to transfer about 250 000 items per year: about 1.5% of total holdings. At this rate it would take 60 years to deal with current holdings. This rate of progress is inadequate because:
- Much of the material will not last for 60 years; average 'format life' of videotape is 20 years or less (as little as 10), and then the format is obsolete. Life expectancy of the material itself varies with storage conditions, but without cold, dry storage most audiovisual materials deteriorate after 20 to 30 years.
- New material comes in; project Presto found that acquisitions were exceeding preservation work by a four to one ratio [5]
- There is already insufficient budget and insufficient resources: the PrestoSpace survey found that archives had half the budget they needed (just for their planned 1.5% per year transfers), and the facilities providers also had half the needed capacity.

The TAPE Survey

A much wider survey was performed by the TAPE [6] project, and after extensive analysis the results were published early in 2008. The survey was very successful: 376 responses, covering collections ranging in size from less than 100 hours to over 1 million hours, and totalling 25 million hours of material. Combined with the PrestoSpace results, and allowing for overlap, the result is specific knowledge of 30 million hours of European audiovisual content: where it is, what it is, its condition, and plans for its preservation.

Principal Results Of The TAPE Survey In general, 70% of material is seen by its curators as in acceptable, good or very good condition, and 30% is deemed deteriorating or unknown. **But half the archives do not have controlled storage conditions, half have no regular equipment maintenance, and 2/3 do not have a systematic preservation programme.** Furthermore, for every response saying that a particular problem (like vinegar syndrome or colour fading) was not present in a film collection, there were three "not known" responses. The strong suspicion is that the 70% figure represents complacency (or apathy), not fact – and that this complacency is a problem equal to, or even greater than, the problems of physical decay and of preservation budgets. Surveys on digital preservation have found similar results: findings that say "we don't see a problem", which is not at all the same as evidence that there is no problem.

Where Is the Material? Earlier surveys had been dominated by broadcast archives, with their massive collections (in the 100k to 1 million hour region). TAPE showed the importance of the very many small collections, the long tail of audiovisual archiving:

- Broadcast archives 30% (roughly)
- National collections 15%
- Other major collections 15%
- Small and specialist collections 40%

These figures, plus the other results, show a situation that is probably common to all collections: a great deal of material is in small collections, where the problems of equipment, technical staff, proper storage and budget are most severe.

The Solution: Mass Digitisation Projects and the *Preservation Factory* Approach

For audio and for videotape, the cost of preservation work can vary enormously. The primary factors controlling cost are:

- Accessing the material: it costs A LOT to search, fetch, and transport the old material, and then shelve the new material. If material is being accessed anyway (for issuing or research), then preservation work can be done at the same time (preservation on demand).
- Throughput: a conventional transfer consists of one person playing an item from an old format onto a new format, and then checking the result. This takes about 3 hours per hour of material, and is the LEAST effective way to transfer a lot of material. A dedicated facility operated as a transfer factory, with each operator running not one but four or five simultaneous transfers, and with maximum automation of checking and labeling (basic metadata) – can save about 2/3 of the cost as compared to a conventional transfer.

The above two points are in serious conflict. Preservation on demand usually needs to be tied to a conventional, one-at-a-time transfer. A 'factory' approach is required to copy a whole collection in a reasonably short time with maximum cost-effectiveness.

The most expensive way ("the obvious way") to do preservation is to access the material specifically for preservation work, and then do the one-at-a-time transfers. The following are the PRESTO rough estimates of the costs of the three options:

Preservation project method	Rough cost per hour for ¼" audiotape, £
"the obvious way"	150
On demand, when the material is coming off the shelf anyway	100
Mass transfers, factory method	50

The problem with the above table is that cost information is notoriously incomplete and imprecise. "Cost" ultimately is what a departmental or institutional budget has to pay – visible cost. If a department has the staff and the equipment and the time, the visible cost for the 'obvious way' may be zero, rather than £150/hr! In contrast, shipping it out the door to an efficient audio transfer factory would incur the very visible £50/hr.

It is important to consider not just cost but time. If the material is already old and causing enough problems to motivate thinking about copying to a new format, then this copying may need to be done in months, not years. The total time as well as the total cost must be calculated for each of the options, and the best choice is the cheapest within the allowed time, not the cheapest overall.

Access and Cost Per Use. The true cost of an asset is total lifecycle cost. The true benefit is related to the number of times that asset is used over the lifecycle. Although not every use has equal benefit, overall more collection access means more benefit. Therefore a simple way to combine transfer cost, life cycle cost, and the significance of new service opportunities, is to translate those new opportunities into a predicted rate of item usage. Options for preservation can then be compared, in monetary terms, on a "cost per use" basis. A significant conclusion of the PRESTO survey is that archive preservation strategy should aim at the "lowest cost per use" over the life cycle of the new media, NOT at the lowest transfer cost.

The Solution for Small Collections The main problem with the preservation factory approach is that it takes a considerable investment to set up and run such a factory. Only institutions with large volumes of audiovisual media could adopt the approach.

One could suggest that the large institutions open their factories to use by smaller institutions. The problem here is that all holders of audiovisual material are in a race against time for preserving their own holdings, and in general do not have spare capacity.

The preservation factory approach is the answer to preservation for broadcast archives, and is suitable for:

- large collections

- that earn money
- and have sophisticated technical support.

What is needed is an audiovisual preservation solution for everybody else.

Preservation is a major issue, but cannot be viewed in isolation. The institutions that hold this endangered material perform services, and broadcast archives serve a highly technical and rapidly changing industry. Preservation strategy needs to consider – and foresee – the future service requirement of multimedia collections for at least the next twenty years. These service requirements will increasingly be based on electronic mass storage and direct, networked end-user access and web technology.

EC project Presto-Space[7] has a solution: fostering the preservation factory approach as commercial services, with archive access as the main method for acquiring funding.

Major audiovisual cultural collections in Europe joined forces to propose Presto-Space, a 6th Framework Integrated Project to provide practical methods for digital preservation of all types of audio-visual collections.

The goal of 'preservation for all collections' requires:
- an integrated approach
- sustainable assets
- easy access
- much wider exploitation and distribution.

The key idea is: an accessible item is more valuable than an item stuck on a she*lf*. An integrated process provides this access, generating revenues that will fund the activity and developing resources to finance collection maintenance. The preservation factory approach delivers the full package: metadata, media, storage, website, asset management. The prospect of much better access is the mechanism for acquiring the needed funding: *access funds preservation*.

Cost and Value

There are three surprising facts about economists and assessment of Public Value:
1. It's what they do. Economists do not generally deal with commercial value – that is for accountants. One of the major functions of economic thinking is to deal with Public Value.
2. Economists measure public value in explicit monetary terms: Pounds Sterling, Euros, dollars and other currency.
3. Economists have been doing this since Adam Smith[8].

Commercial and Public Value Archives need money in order to maintain their contents and continue to provide access – which is their function. The money comes from various sources, but one possibility is commercial sales. These sales have a *value* that is obvious and easily quantified. The problem with obvious things is that they can force less obvious issues into the background, even if the less obvious consideration is, really, the more significant.

Figure 2- A photo of a house – or a mountain?[9]

There is a general idea that audiovisual archive content has to be protected from free access for two main reasons:
- There are underlying rights that have not been cleared
- Free access will destroy the possibility of commercial sales

These positions are assumptions, not reasons. With regard to underlying rights, the assumption is that the rights-holders do not want public access. This assumption is often stated in virtually the same breath as a statement about the difficulty or impossibility of finding rights holders. Obviously the rights holders have not, in general, been consulted. However there are many examples of rights holders who would like nothing better than free access, or indeed any access, because a principal problem with many collections is that access is so very limited

An emerging phenomenon is rights owners – creators of musical or literary works – who give their works away on MySpace[10] or use other Internet-based mechanisms, in order to develop a following – or simply to be heard and read.

Examples are:
- "The Arctic Monkeys came to prominence through the MySpace site, and Lily Allen was signed to a records label after one of her songs proved popular on there."[11]
- Five collections of stories and novels by Cory Doctorow are available free from his website[12] "*enlisting my readers as evangelists for my work and giving them free ebooks to distribute sells more books. As Tim O'Reilly says, my problem isn't piracy, it's obscurity. Best of all, giving away ebooks gives me lots of key insights into how to make money without restricting the copying of bits. It's a win-win situation*".[13]

With regard to protecting commercial value, the data from one major broadcast archive is that, over the 70-plus year life of the archive, only 5% of the content has been exploited commercially. The consequence is that 95% of content is restricted, for the sake of the 5% that could gain commercial exploitation. This situation is definitely a case of seeing the house and not the mountain, in .

In the work of economists who do quantify public value, the situation with the archive just described is more like "only 5% has realised one form of monetary value, and that has prevented the rest from realising another form of monetary value which could be very much greater". Under this second interpretation, it is more like "we sold the house, because we didn't know we could make a national park out of the mountain, and spawn a range activities with a huge total monetary value."

Pollock[14]gives the following examples: "The New York Times and The Guardian both give free access to their online edition but feature advertising as well as selling a corresponding print edition (complementary goods). … And Yagan (2005) recounts his extremely successfully experience of 'giving away' content in the form of online textbooks first at SparkNotes and then, subsequent to its acquisition, at Barnes and Noble."

The conclusion is that realising *public value* should become the major consideration for audiovisual collections. The next section looks at methods for releasing that value.

Releasing value

Commercial value is based on selling goods and services. The issue with regard to archives is the changing nature of the market. Anthony Lilley refers to the conventional market as the "age of scarcity" – with media industries concentrating on a limited range of high-turnover, big market products. Lilley describes "the skills of the age of scarcity – hitmaking, hype and cross-media promotion among others." [Ref11, p7].

This is a marketplace of restriction and protection:
- restricted range of products that have to 'earn their shelf space'
- restricted access: come to our theatre or cinema, buy our recording or publication, tune in to our channel – because we are 'sole source'
- restricted rights: look but don't touch, passive viewing and listening but no active involvement. Difficult to 'quote' media even for academic purposes that are allowed under copyright legislation; illegal and very difficult to make copies or extracts for general purposes.

The new marketplace created by computers and the Internet is altogether different. This market has no shelves, it is as easy to stock 100,000 items as to stock 100 items, and the major issue is 'presence' (standing, trust, brand recognition). This is a market requiring the "skills of the age of plenty, the exploitation of the so-called long-tail, the importance of authenticity of voice, and the power of communities of interest."[Ref 11, p7]

In these new markets, the goods which previously couldn't get on the shelves are transformed into major sources of profit. The whole 'long tail' phenomenon is a recognition that there is more money to be made for 'worst sellers' than from 'best sellers'. The 'worst sellers' are at the low frequency-of-sales long tail of the distribution of items (according to sales rate). Amazon the online

bookstore (and now books-and-everything-store) is often sited as making more money from the long tail than from the popular goods[15], and EBay has achieved its position (profits of US$310m on turnover of US$1.4bn for one quarter – Q2 – in 2006)[16] by offering a way to sell the longest tail of all: all of our used goods – a market that had been considered insignificant.

These markets work by attracting people to 'use' the site, a new combination of goods and services. It could be compared with the "come to our national park" approach to commercial sales, where people do come to use the park, but they also buy things.

The situation is complex, because the most successful new company of all, Google, doesn't sell goods or services. It provides a free service, which attracts users. The users come to Google as an intermediary point in their quest for information (in the widest sense) in this new 'activity place' called the Internet. Google then sells pinpoint advertising space, tuned to the wants of the users. The traditional corner shop had an element of this model, when it charged 20p for people to put cards in their front window. But every visitor to the shop had to search all the cards to see if any were of interest. The Google Corner Shop waits for a customer to ask the shopkeeper for a magazine about cars – and then all the cards selling cars or car repairs come flying out of the shop window and hover about the customer.

The point for audiovisual collections is: people like audiovisual content. If audiovisual collections were online, people would come to the Audiovisual Archive Shop, to look around – and if enough people come then that creates the kind of opportunity that Google has so effectively exploited. The point about the long tail is that in an online world there is no such thing as a 'worst seller'. The obscure or specialist collection can have its day.

For the obscure archive to 'have its day' via the Internet, it must be:
- **online** – this requires digitisation, encoding and hosting
- **public** – because any restriction on access is a proportionate restriction of interest, visitors to the site – and ultimately any value, public or commercial, associated with the site
- **visible** – the hardest condition of all. As in the Cory Doctorow quote [Ref **Error! Bookmark not defined.**] "my problem isn't piracy, it's obscurity". Ways to become visible are a huge subject, but the above two points, online and public, are preconditions.

Public value is released in a very simple way – the material is simply opened to public access. The commercial market requires skills of merchandising and mechanisms of brand building, including a great reliance on advertising. Public value is much more straightforward: it merely requires removal of restrictions.

Commercial and public value differ diametrically in their approach to access. For commercial markets, access is restricted – to release commercial value. Restriction may be entirely necessary for manufactured goods and 'commodities'; for the new information market it can become a lose-lose-lose situation.
- The economy gets much less benefit than if public value were also realised;
- The general public get restricted access;

- Rights owners do not enter the new markets (those based on internet access and the 'long tail'), and generally lose out as the old markets dry up.

There is a particular danger for heritage material coming, originally, from the commercial sector. This was dramatically illustrated by the fate of cinema films in the last century:

"For the first few decades of the 20th Century there were few official audiovisual collections. Film stayed with the companies that produced it. Many such companies lasted only a few years. Their accumulated negatives were seen as 'used goods' of small value, and so roughly 2/3 of the film production pre-1940 no longer exists"[17]

The example shows the disastrous consequences of allowing commercial value to dominate public value: short-term commercial value is at odds with long-term heritage considerations. The 'restrict, then dump' commercial drivers end up opposed to both preservation and access.

Where commercial value depends upon restricted access, public value requires enlarged access, which can then produce a win-win-win situation.
- Economic benefit is maximised, certainly for media which is part of the 'information industry', where the public value dominates (if allowed to!);
- The general public get enlarged access;
- With the enlarged access, rights owners can, if they wish, develop new business models that can be far more remunerative – in commercial value – than is the case for the old models.

Audiovisual Digital Repositories

Audiovisual archives, and indeed all who depend upon digital storage, should have planned from the beginning for how digital technology would be used to fulfill storage and access requirements in a sustainable fashion. However over the last 15 years (from roughly 1995) the rate of technological change into digital audiovisual media has been so fast, and the pressures for audiovisual preservation of analogue media have been so severe, that audiovisual collections have concentrated on migrating into the digital world, and are only now really beginning to ask – in detail – how they can continue to live, and live well, in this new world.

Storage   It seems obvious that if material is to be digitised, then it needs somewhere to reside – something to 'hold the bits' – and so a repository it an implicit requirement. Actually, many types of digitisation – in the exact sense of the representation of content in a digital fashion – produce hold-it-in-your-hand media such as CD, DVD and data tape, and do *not* need a repository

These *discrete media* do a lot:
- they hold the results of a digitisation process
- they can hold such results for some number of years or even small number of decades.

However discrete digital media provide no better access than the current situation of analogue tapes on shelves. An essential characteristic of a repository is some form of digital mass storage, for the specific purpose of **access.** If electronic access were not the determining factor, a repository would not be required and archives would simply fill their shelves with discrete digital media.

Access   The immediate reason for a repository (which hasn't been defined, but a picture is forming) is to provide an improved alternative to media on shelves, in order to take advantage of electronic methods of access.

Clearly electronic access offers obvious improvements:
- speed of delivery
- elimination of contention for copies
- elimination of circulation control: booking, tracking, overdues

Not so obvious are the new methods for access, and new reasons for access. These are not so obvious because they may not yet exist, or be at an early, primitive stage.

There are two main parts of digital access:
- electronic delivery
- web access

Electronic delivery

Audiovisual archives have been notoriously inaccessible, because the contents of the archive were on professional audiovisual formats that required careful handing and expensive equipment for playback. Electronic delivery of audiovisual content solves a number of problems simultaneously:

Contention for copies: elimination of the 'somebody already has it' problem, and elimination of the need to take steps to avoid that problem, such as making multiple viewing copies

Circulation of control: checking out, delivering, and ensuring the return of physical media. However elimination of physical circulation could result in elimination of all knowledge of who has accessed material, so a new sort of circulation control is a requirement of a digital repository

Physical attendance at the archive: electronic delivery can go anywhere (given adequate bandwidth). However if people are no longer presenting themselves within the walls of an audiovisual collection, the new problem is introduced of how to know who is entitled to access the material. Indeed, the whole restriction of access to 'legitimate research', which is built into the rules and laws of many collections, is called into question by electronic delivery.

Damage to the material. Essentially electronic delivery makes 'viewing copies on demand', so the 'original' – however that is to be defined in a digital environment – is unaffected by access. This statement needs qualification, because access to digital files on optical or magnetic media does potentially involve wear – although in practice it may prove far better to read files, and check error rates, than to leave files unread an not know about physical deterioration within the repository. Indeed digital repositories may be no more immune from media wear than conventional tapes on shelves – it is just that they should have automated, efficient, cheap and reliable methods for checking and regenerating data.

In short, electronic delivery removes *all* the physical, technical and logistical barriers to *unlimited* access to audiovisual collections. The effect will be to make the legal barriers all the more obvious – they will be all that's left. Inevitably, all the pressure to hear and see audiovisual material – from a world audience – will be applied specifically to these legal barriers. Given the amount of change to the physical barriers in just the last few years, it would be reasonable to expect that the next one or two decades will see significant changes to the legal situation. The implication for repositories is that their electronic delivery

capabilities may have to grow far beyond the current access requirements for today's audiovisual archives.

Web access   Web access is also electronic – a data connection between a user and the repository. The reason for separating web access is that it is primarily concerned with finding material, whereas electronic delivery is about what happens after material has been found.

The web needs highlighting because of several factors. The web:

- is in people's homes (in Europe)
- has come to define what does and does not exist
- has spawned a variety of search mechanisms
- is becoming the preferred source for at least some audiovisual material, beginning with commercial music tracks.

The combined result of these factors is as follows: audiovisual collections will need to be 'on the web' to show that they exist; users from across the world will find them; these users will expect access to their contents.

An audiovisual collection thus is in a difficult position: if not on the web, it effectively doesn't exist. If on the web, it raises an expectation that, largely, it cannot fulfil. **Meeting the *expectations of the web* is likely to be the principal challenge of the next decade, for audiovisual collections.** It follows that providing for web access, in terms of 'discovery metadata', the organisation, quality and standardisation of that metadata, and in terms of delivery of actual content at various qualities and bandwidths, will be principal requirements of the repository.

Sustainability   The main issue for a repository of archive material is that *the content has to outlast the technology that carries the content*. This is not a new issue, and audiovisual archives have been migrating analogue content for decades. It is the digital technology that has little experience of migration, and most of that is pretty uncomfortable experience.

As an example, databases have migrated from mainframes to mini-computers, and from one commercial product to another. Libraries have about 30 years of such experience (migrating library catalogues), and every migration is seen as a major difficulty, with much complaint for years afterwards about what was or wasn't lost in the migration.

The problem is that functionality is not easily and fully migrated from one IT system to another, where system now has to include desktop operating systems, network technology, middleware of various sorts, and finally the actual application (ie library catalogue) involved. New IT technology inevitably has new functionality, different ways of handling old functions – and insidious ways of making some things impossible.

What has this to do with audio and video? The audio and video will likely migrate very well as IT technology changes – at least, the master, full-quality (uncompressed) audio and video files should move with little effort, as compared with software evolution. The issue is in the functionality that defines the repository: the functionality that manages the audio and video. Such functionality is not remarkably different from our current conception of library IT systems, and the 30-year history of

migration of such systems has been one of major effort and major problems.

A repository will be a fairly complex IT system, whose functionality is meant to survive intact, over many decades, despite the obsolescence and replacement of **every** component of that functionality. The key to sustainability of an audiovisual collection is a problem common to all "digital library" collections: the sustainability of the functionality of the repository – not the sustainability of the audiovisual content itself. In simpler terms, the audio and video files can be expected to last with straightforward 'data refreshing' technology; it is the system providing access to those files which presents the large risk.

Lessons for All Archives

- Use a factory approach for cost effectiveness
- Mass storage provides access as well as preservation
- Storage is a service
- Storage costs are becoming negligible
- A digital archive is a treadmill, requiring maintenance
- Access provides funding
- "Our problem is not piracy – our problem is obscurity"

References

[1]   SCHUELLER, D: Making Audiovisual Contents Available, EuroChina 2002; www.eurochina2002.com/docs/Online%20after%2024%20May%202002/Conference_Presentations/Dietrich_Schuller.pdf
[2]   http://presto.joanneum.ac.at/
[3]   FOCAL ref
[4]   PRESTOSPACE ref
[5]   http://presto.joanneum.ac.at/projects.asp#d2
[6]   TAPE
[7]   www.prestospace.eu
[8]   en.wikipedia.org/wiki/Adam_Smith
[9]   Stodertal Valley, Flickr, Creative Commons licence; http://www.flickr.com/photos/swq/398764643/
[10] www.myspace.com
[11] Inside the Creative Industries: Copyright on the ground. Anthony Lilley IPPR July 2006 www.ippr.org/publicationsandreports/publication.asp?id=483
[12] www.craphound.com/novels.php
[13] http://www.craphound.com/bio.php
[14] The Value of the Public Domain, Rufus Pollock IPPR July 2006 www.ippr.org/publicationsandreports/publication.asp?id=482
[15] The Long Tail, Chris Anderson. Wired 12.10, Oct 2004 www.wired.com/wired/archive/12.10/tail.html
[16] www.theregister.co.uk/2006/07/20/ebay_Q2_2006/
[17] Film Preservation Study, National Film Preservation Board, 1993. http://lcweb.loc.gov/film/filmpres.html

Author Biography

Richard Wright trained at the University of Michigan, USA and at Southampton University, UK. He has worked in acoustics, speech and signal processing for US and UK Government research laboratories, University College London, the Royal National Institute for Deaf People and Cirrus Research. He headed the EBU panel on Future Radio Archives and EC project PRESTO, and has been Technology Manager, BBC Archives since 1994.