Structural requirements for digital audiovisual preservation

Tools and Trends -- Koninklijke Bibliotheek

The Hague, 1-2 November 2007

Richard Wright, Archive Technology Manager, BBC, London UK

richard.wright@bbc.co.uk

Overview

1. What's Special About Audiovisual?
2. Preservation
3. Digitisation
4. Storage
5. Access
6. Cost and Funding

1- What's Special About Audiovisual?

- A lot of it
- Intrinsic interest
- Intrinsic problems!

A lot of it

- 200 million hours
  - In archives!
- 30 million hours in collections documented by PrestoSpace and TAPE projects

TAPE survey

Type of organization	No.of resp.	%
Archive	143	38.03%
Library	81	21.54%
Museum	42	11.17%
research institute	28	7.45%
Institute	26	6.91%
radio/tv company	21	5.59%
Other	15	3.99%
commercial company	11	2.93%
private person	9	2.39%
	376	100.00%

TAPE Survey

Holdings	Total responses	Total amount in hours
Film	221	3 049 667
Audio	327	10 878 904
Video	313	11 942 578
Total	861	25 871 149

Intrinsic interest

- Walking, talking record of the 20th Century
- Film and Broadcasting: popular media
- JISC ‘most desired digital library content’

Intrinsic problems!

- Access has always been restricted
- 75% of current content at risk
  - Fragile
  - Decaying
  - Obsolete
- Only 30% within current ‘preservation planning’
  - And only half of that is funded

2- Preservation

- Digitisation seen as an answer
- Analogue conservation is not an option (except for film)
- Emulation is not an option
- So: high-efficiency migration
  - And permanent migration = maintenance

3- Digitisation

- Carrier vs Content
  - Abandonment of the original
  - Content set free from the carrier
    - Same for digitised books?
- Issues:
  - File formats, codecs, wrappers, and necessity to avoid “lossy compression”
    - Same for other ‘files’?

BBC Preservation Requirements

- Moving about half the analogue holdings into digital formats
  - 2”, 1”, U-Matic, Ekta, SepMag, vinyl/shellac, ¼” tape
- Moving ‘digital audiovisual media’ into standard IT formats (datatape, disc)
  - DAT, DVD, D3
- BBC material only
- £50M, 10-years, ends 2010
- Needs a follow-on project of similar size!

BBC Digital Preservation

First problem: what is an ‘archive’

- **Where data lives**, vs where data goes to die, and has to be ‘restored’
- **Where people go when they want something**, vs where data goes when it is no longer wanted
- **An application**, vs the backup or overflow behind an application

Digital Preservation Requirements

- ***Persistence***
  - ***3% loss after 20 years***
  - ***1% as 'realistic target'***
- ***Currency***
    - ***retrieval of files that are usable***
    - ***format, encoding, carrier, file mgmt, OS …***
    - ***implication: 5-year migration cycle***
    - ***"100-year" storage media of slight interest***

www.ebu.ch/en/technical/trev/trev_308-archives.pdf

Digital Preservation Req's (2)

- ***Wanted: cost curve for % loss over 20 years** (actuarial statistics)*
  - ***Probability of loss***
  - ***Distribution of the probability of loss – what is the size of the uncertainty?***
  - ***Cost function: how does probability of loss vary against increase – or decrease – in investment?***
  - ***Variation over time – again, basic information for investment decisions***

Digital Preservation Req's (3)

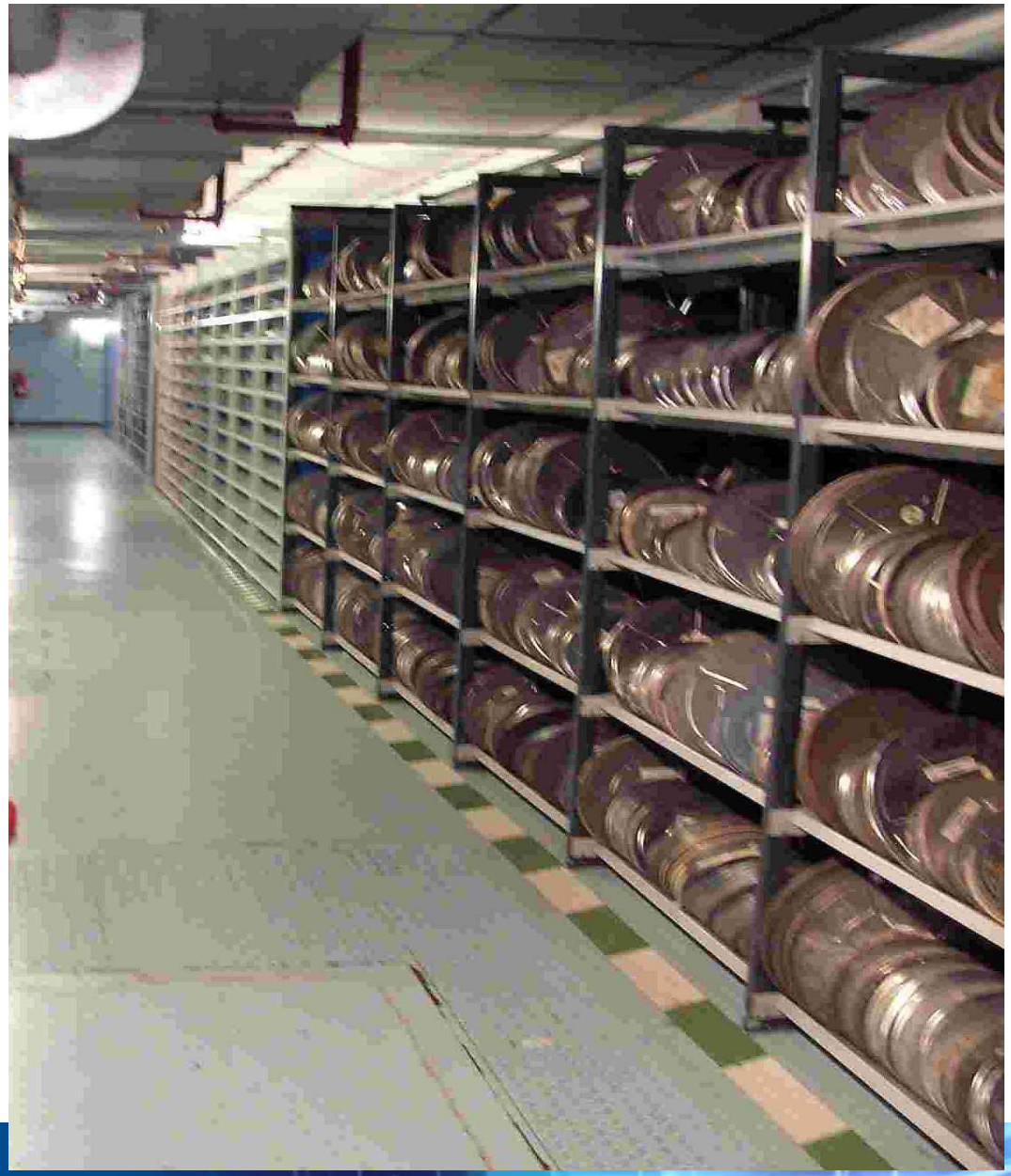
- ***Wanted: error-tolerant storage***
  - ***Equivalent of 'line dropout concealment' that we've had with videotape (analogue or digital)***
  - ***Distribution of read-errors across files, reducing SNR rather than failing or making a distinct gap***

Current work in DTI projects Estastar and Avatar

4- Storage

- Archive media
  - Longevity of ‘briefest link’
- Mass storage
  - Total lifecycle cost
  - Including next migration
  - Detailed modelling
- Storage management

BBC Archive: Storage requirements



100 km of shelves

650k hours video

350k hours audio

2M stills; ¼ online

3M items sheet music

400k “pronunciations”

1.5M titles in “grams library”

BBC business documents

95% internal use

What the Archive Does

BBC “Media-Handling” Requirements

- Business – what we do
- Storage – local and archive storage
- Distribution – broadcast, Internet …
- Preservation – analogue to digital
- Digital Preservation – and then what?

Business requirements

- **Transmitted Material**
- **Raw Material = Rushes**
- **Archive Material**
  - **Use of the Archive**
  - **Preservation of the Archive**

Transmitted Material

- **TV: 1000 hrs/week**
  - **UK national output; not counting commercial channels and abroad**
- **National Radio: 1500 hrs/week**
  - **UK national output; not counting 'nations and regions'**
  - **Excluding repeat!**
- **World Service: 1000 hrs/week**
  - **In over 30 languages**
  - **May include repeats, translations**

Raw Material = Rushes

- **Shoot to Show ratio**
  - ***as high as 40 for high-end productions***
- ***News: 300 hrs/day in and 15 hrs/day out***
- ***Up to 15 000 people***
- ***Overall 10x 'shoot to show' =>***
  - ***10 000 hours per week of video***
  - ***Doesn't include multiple internal copies!***

Archive Material

- ***Use of the Archive:***
  - ***3000 enquiries / week***
  - ***10,000 items / week***
    - ***Half the video in 'browse quality'***
    - ***Just starting to deliver audio files***
- ***Preservation of the Archive***
  - ***2000 audio items/week***
  - ***1500 video items/week***

Storage Requirements

- **Transmitted Material: 1000 hrs/week**
  - **Standard Definition: 200 M b/s**
    - **$\Rightarrow$100 GB/hr => 100 TB/week**
  - **Moves to 400 TB/week with conversion to High Definition**
- **Raw Material: 10k hours/week**
  - **1 PB/week, moves to 4 PB/week**

Storage Requirements II

- **Archive Material**

  Intake is about 300 hours/week = 30 TB per week, moving to 120 TB/week with move to HD

  Preservation: 1500 video items per week is about 800 hours/week

  => 80 TB week

  which will NOT increase with move to HD, as we will not 'upconvert'

Distribution Requirements

- Transmission of 8 concurrent streams at 4-6 Mb/s each
  - Real problem: simultaneous *analogue switch-off* and change to HD
- Web “transmission”: could be millions of simultaneous requests, so BBC iPlayer will use peer-to-peer distribution
  - 4.6M hrs of requests in 2005/6; 60% incr on 04/5
  - Aside: for web, transmission costs swamp the storage costs

5- Access

- Historically a problem
  - "Finding Murphy Brown" Jeff Ubois http://journals.tdl.org/jodi/article/view/jodi-177/155
- New culture: open everything
  - Outside AV archives – software, research, books, music, images …
  - Within AV archives
    - Archivespourtous
    - Prelinger model
    - BBC rediscovers public value => Open Archive
      - **www.bbc.co.uk/archive**
  - YouTube

6- Cost and Funding

- Digitisation isn't interesting – Access is!
- Combining preservation with greatly enhanced access is interesting
- Delivering enormous increases in usage is interesting
- Access and usage can be translated into funding
  - Using public value

Analysis: Digital Preservation- can we afford it?

- S Chapman, Harvard Lib: “Is Repository Storage Affordable?” Jr Dig Inf, v4 iss2
- For one 333-page book, per year:
  - Book on shelf: $0.31 (standard vault)
  - Microfilm in film vault: $0.19
  - Scanned images in OCLC repository: $0.47 (at $15/GB)

Audiovisual Digital Preservation costs

- Chapman: Harvard book repository vs OCLC digital repository
  - Shelf costs: $10/ft => $0.10/GB for videotape (including the media cost)
  - Repository: OCLC: $15/GB
- J Palm, Swedish Natl Arch: “The Digital Black Hole” www.tape-online.net/docs/Palm_Black_Hole.pdf
  - “One system manager per terabyte” => $300k / TB
  - Best: One system mgr/10 TB => $30k / TB = $30/GB

Affordable Digital Preservation

Can a *repository* be anything other than high-end managed servers/discs?

- Either more efficient online storage
  - But: needs to aim for $0.10/GB
- Or: digital data nearline, or on shelves

Managed Nearline/Offline Storage

- Key Issue for digital preservation
- Needs to
  - Have low labour costs
  - achieve 99% *persistence* over 20 years
  - Observe *trusted digital repository* standards => OAIS for shelf-based media (and for tape robots)

Storage Costs Forecast

Cost per GB for video

(Managed Online has two costs: media+management)

Year	Managed Shelves	Mg'd Online	Un-Mg'd Offline Disc	
2002	$0.10	$15=8+7	$4	
2006	$0.11	$11=4+7	$1	
2010	$0.12	$9 =2+7	$0.25	
2020	$0.15	$7 =0+7	$0.02	$0.02

Key issue: **cost of managing** offline storage

(because unmanaged storage IS a black hole)

AV Preservation Strategy 1: Analogue

Analogue Preservation of 1 hr full quality SD video, for 10 years

- Two copies on shelves = $30
- One transfer every 20 years @ $160
  - 40% inflation over 10 years => 0.7
- => $142
  - BUT: totally impractical owing to format obsolescence

AV Preservation Strategy 2: Digital

Digital Preservation of 1 hr full quality SD video, for 10 years

- Digital media storage: $10 at start, $2.50 five years later
- + Datatape copying: $16 (=$160/10)
- + two copies on shelves = $30
- Total = 2x$10 + 2x$2.50 + 4x$16 + $30
- => $119 [not $6000 as per Chapman]
- Datatape on shelves – or possibly USB disc drives on shelves – or a tape robot
- Using OAIS standard – offline!

A Call to Arms

'My problem isn't piracy, it's obscurity'
Tim O'Reilly, O'Reilly Press

- Archives of the world: give your contents away – online!
- You have nothing to lose but your obscurity

  and low interest, low usage, low budgets – and all that goes with them

Cost vs Value

- Only 5% of BBC archive has been commercially exploited
- The public value has been ignored – until now: Creative Archive, iPlayer, Open Archive
- Dutch Archive has received €173M for preserving audiovisual heritage, and making it available for public use
- Equivalent UK figure: €642M

How much to ask for

Country	Bid, € millions
Austria	88
Denmark	58
France	606
Germany	805
Ireland	55
Italy	545
Netherlands	173
Portugal	70
Spain	361
Sweden	93
U.K.	642

Exploitation strategy and costs

- Use national portals, European Archive, VideoActive, You Tube, Flickr, TEL, EDL ---
- Give away everything possible
- Become *indispensably cool* for students
  - All of which costs nothing
- As much content as possible on website – “the government wants the people to see it!”
- Partnerships with rights holders – ask them for money! -- “the government has paid its share, now you pay your share”

Thank You

- PrestoSpace
  - www.prestospace.eu

  - richard.wright@bbc.co.uk