The Digital Preservation of Audiovisual Content – or – what's so special about audiovisual?

Richard Wright, PrestoPRIME

BBC Research & Development, UK

SUN Preservation and Archiving Special Interest Group

Malta 24-26 June 2009

Stone, papyrus, film, hard drives

Medium	bits/cm²	life
Stone	10	10 000
Paper	10^4	1000
Film	10^7	100
Disc	10^{10}	10

⇒Each change 1000 times more dense and hence 1000 times *cheaper*

but the media lasts 0.1 times as long

⇒Interval between changes also 0.1 times as long

The problems of AV content

- Traditional media are obsolete (except film, and that's "endangered")
- So we've been digitising
  - Probably 5 to 10 million hours of digitisation by broadcasters in last decade, in Europe
  - National audiovisual collections add millions more
  - Born digital: archiving 300 hrs/week = 30 TB (BBC)
    - Uncompressed Standard Definition; HD will be ???
- So: moving from media that lasts a few decades, to files – and to technology that lasts a few years
- So we'd like to use digital library technology
- If we could

PrestoPRIME: problems of using digital library technology in broadcasting

- Because MXF is the primary professional broadcast wrapper format
  - In Europe, at LOC, in US Public Broadcasting and for digital cinema
- Because MXF isn't supported by any digital library tools: JHOVE, PRONOM, metadata extractors
- Because OAIS is little-known in broadcasting
- Because broadcasters use MAM, not digital archives much less digital repositories
- Because AIP etc has few broadcast exemplars
- Hence: a shotgun marriage of OAIS and MXF

The technical registry
PRONOM

Search Results

Simple search | File format | PRONOM Unique Identifier | Software | Vendor | Lifecycles

You searched for: "mxf"

Save as... XML | CSV Print

No data matched the search criteria you entered.

The National Archives, Kew, Richmond, Surrey, TW9 4DU. Tel: +44 (0) 20 8876 3444. Contact us

Search Results

? Help : report on file format

Simple search | File format | PRONOM Unique Identifier | Software | Vendor | Lifecycles

You searched for: "wav"

Save as... XML | CSV Print

page 1

PRONOM Unique ID	Format Name	Format Version	Extension	Format Risk
fmt/1	Broadcast WAVE	0	wav	
fmt/2	Broadcast WAVE	1	wav	
x-fmt/397	Exchangeable Image File Format (Audio)	2.0	wav	
x-fmt/389	Exchangeable Image File Format (Audio)	2.1	wav	
x-fmt/396	Exchangeable Image File Format (Audio)	2.2	wav	
fmt/6	Waveform Audio		wav	
fmt/141	Waveform Audio (PCMWAVEFORMAT)		wav wave	
fmt/142	Waveform Audio (WAVEFORMATEX)		wav wave	
fmt/143	Waveform Audio (WAVEFORMATEXTENSIBLE)		wav wave	

Page: 1 of 1

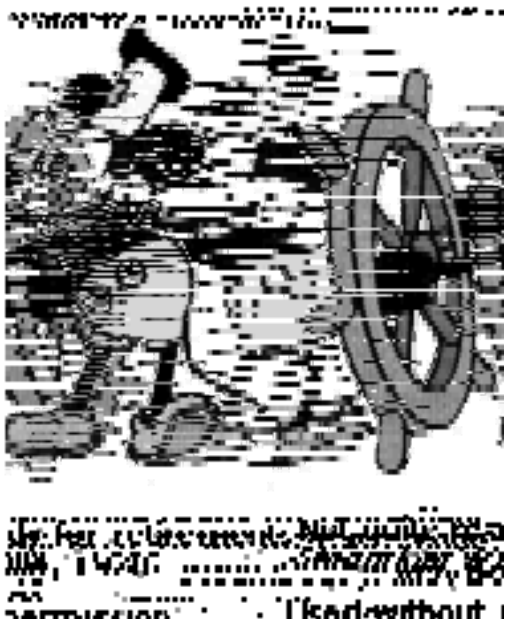
page 1

Policy: Compressed vs Uncompressed

- Why not go from one high-quality compressed format to the next, forever?
  - Because there will not, in general, be a transcoder
  - So the material c1 has to first be decompressed to u1, then encoded to c2
  - At the next cycle, c2 is decoded to u2
  - u1 ≠ u2 (≠ u3 ≠ u4 …)
  - So the “real thing” drifts away
- And- an issue of file resilience (mitigation of loss):

Mitigation of Loss

BMP no errors 40k



GIF 3 errors in 10k

BMP 160 errors in 40k

Heydegger, V. (2008). Analysing the impact of file formats on data integrity. *Proceedings of Archiving 2008*, Bern, Switzerland, June 24-27.

Format Roadmap: low quality media

Ingest Format	Migration format	Notes
VHS tape	DVD	**Access** Perfectly adequate for VHS playback
VHS tape	MPEG-4 files	**Access** Adequate for quality. Minimum data rates (MPEG-4): 500k b/s. There are MANY potential access formats, and they come and go.
VHS tape	DV files	**Archive** (temporary) 25 M b/s, 12 GB/hr. Migrate to lossless for preservation.
'low end' digital files	Save as is, AND save as DV or lossless	**Archive** (temporary) Before format or DV format becomes obsolete, migrate to lossless for preservation.
DVD	DV files	**Archive** (temporary) 25 M b/s, 12 GB/hr. Migrate to lossless for preservation.

Conclusions about compression

- **Should only be used once** in a ‘life history’ of audiovisual content
- Then have to move to uncompressed to avoid the ‘drift’ of never having a fixed reference
- But: compressed is less resilient
- So: only use interframe compression (as in DV and high-level MPEG profiles)
  - Which would allow individual frames to be lost without affecting the rest of the file
  - If only partial recovery of files were possible!

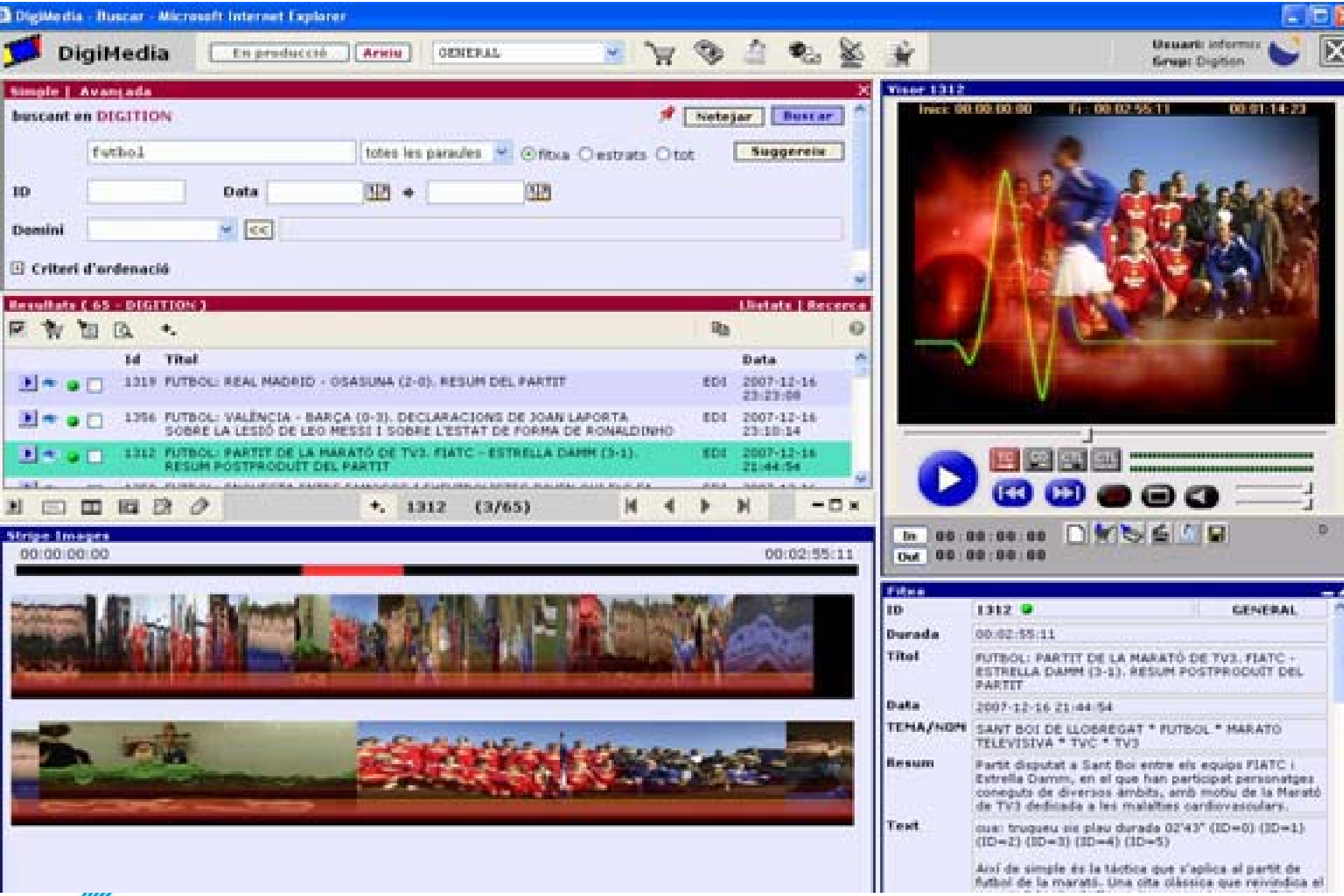
Strategy: Migration, Emulation

- Migrate to uncompressed as soon as you can
- Check every format for risks, obsolescence, every N years
- Keep the original bitstream – if you can!
- Can use emulation to extend the usable life on an at-risk format
- Never move from lossy A to lossy B
  - Possible special pleading for genuine transcoding
- Implies possibility of keeping a poor format until it is obsolete – then migrating once to a not-so-poor format – then migrating to uncompressed
- Lossless compression (interframe compression only) is nearly as robust as uncompressed, for 1/3 the storage (but adds complexity and dependencies). Total risk unknown!


The Preservation Game
START HERE
Is your AV content in a file format?
NO
YES
Transfer it to a file
Is the format at risk?
YES
NO
Can you afford to store it uncompressed or lossless?
NO
YES (Uncompressed)
YES (Lossless)
Transcode to a lossy format (choose something well supported)
Wait 5 years
Decode to a lossless format
Store files temporarily
Encode to a lossless format
Store files forever
There is no such thing as a safe format, e.g. the idea of 'files' will one day be obsolete
FINISH HERE

A Brief Reminder of Time

- AV content is time based (you knew that)
- And the files are very large (you knew that too)
- So: important for users to have a description of the internal structure of a file – so the user can navigate to ‘the interesting bit’ eg storyboard, “stripe-image”
- Requires: time-based metadata
  - Time-based presentation of the file’s contents
  - An ability to stream / download just ‘the interesting bit’
- Building communities: requires time-based *citation* and *annotation* of content (and sharing them)


DigiMedia - Buscar - Microsoft Internet Explorer
DigiMedia
En producció
Arxiu
GENERAL
Simple | Avançada
buscant en DIGITION
futbol
totes les paraules
Suggereix
ID
Data
Domini
Criteri d'ordenació
Resultats (65 - DIGITION)
Llistats | Recerca
Id
Títol
Data
1319 FUTBOL: REAL MADRID - OSASUNA (2-0). RESUM DEL PARTIT
1356 FUTBOL: VALÈNCIA - BARÇA (0-3). DECLARACIONS DE JOAN LAPORTA SOBRE LA LESIÓ DE LEO MESSI I SOBRE L'ESTAT DE FORMA DE RONALDINHO
1312 FUTBOL: PARTIT DE LA MARATÓ DE TV3. FIATC - ESTRELLA DAMM (3-1). RESUM POSTPRODUÏT DEL PARTIT
1312 (3/65)
Visor 1312
In
Out
Strips Imagen
00:00:00:00
00:02:55:11
Fitxa
ID
1312
GENERAL
Durada
Títol
FUTBOL: PARTIT DE LA MARATÓ DE TV3. FIATC - ESTRELLA DAMM (3-1). RESUM POSTPRODUÏT DEL PARTIT
Data
TEMA/NOM
SANT BOI DE LLOBREGAT * FUTBOL * MARATO TELEVISIVA * TVC * TV3
Resum
Partit disputat a Sant Boi entre els equips FIATC i Estrella Damm, en el que han participat personatges coneguts de diversos àmbits, amb motiu de la Marató de TV3 dedicada a les malalties cardiovasculars.
Text

Access work in PrestoPRIME

- Expressing legacy metadata in RDF
- Combining legacy ontologies
  - And using the result to support search (invisibly)
- And time-based annotation tools
  - Trying to follow new W3C work
- And applying access technology to Europeana, the European Digital Library
  - (plus an effort to quantify, codify and automate rights management – for broadcast content)

Other things I haven't said about PrestoPRIME:

- It's an EC Integrated Project
- Running from Jan 2009 for 42 months
- It's the only project in its 'cohort' about broadcast content
- It's about digital libraries, rights, access
- Ex Libris an important partner: real-world contact with digital library technology
- But also delivering open source implementations of tools developed by PrestoPRIME
- Launching a Networked Competence Centre to live on after the project, to support AV preservation

Thank you

- PrestoPRIME
  - INA, RAI, BBC, B&G, ORF
  - Univ's of Amsterdam, Innsbruck, Liverpool, Southampton
  - Europeana, Joanneum Research
  - Ex Libris, Eurix, Doremi, Technicolor

- www.prestoprime.eu
  - (but not much there yet!)

richard.wright@bbc.co.uk