UCD - Caring for digitised and born digital AV objects.

kieran.o.leary@gmail.com

About me

- Access and Digital Collections Developer in Irish Film Institute
- AV file format advice via Digital Repository of Ireland
- Co-maintainer of AMIAopensource project: ffmprovisr
- Open source software contributor
- Freelance, for example California Revealed Project with U.C Berkeley

Itinerary

- Introduction to video
- Technical Metadata via mediainfo
- Vendors
- Quality Control via QCTools and MediaInfo
- File format normalisation
- File formats

Focus on digital files

- Why focus specifically on digital AV files today in this workshop?
  - Film - requires specialised storage, expertise, equipment.
    - Consider depositing film in a film archive, eg the IFI Irish Film Archive
    - Without specialised equipment, digitisation may be your only viewing method.
  - Videotape - All magnetic AV formats are obsolete.
    - Video also requires specialised storage, expertise and equipment.
    - Videotapes require migration to digital files ASAP as all magnetic AV is already obsolete!

Scope

- We will only cover AV-specific topics that can supplement your existing collections management
- General digital preservation principles will not be covered (not enough time!)
- AV basics
- Metadata tools
- QC Tools
- You will not be an expert, but you will know a little more!

Video

- Video is an ambiguous term, but generally, it refers to the recording or playback of audio and/or images using analog or digital signals.
  - *“Video denotes an analogue or digital signal with audiovisual content that has to be interpreted by playback equipment or software in order to be reproduced.” - Memoriav (http://memoriav.ch/wp-content/uploads/2017/09/Empfehlungen_Digitale-Archivierung_Version1.1_EN.pdf)*
- Difference between Analog and Digital:
  - Analog uses continuous electrical signals to store information
  - Digital uses numbers (ones and zeroes) to store information

AV - Challenges

Question:

What kinds of challenges does AV preservation present?

AV - Challenges

- Large file size
- Complex objects (Complex packages)
- Complex environments/requirements
- Hardware intensive (Uncompressed/ > SD)
- Need for specific metadata standards (Your existing cataloguing system may need to be supplemented with PBCore/EBUCore/EN15907 etc..)
- Requires ongoing condition and obsolescence checks
- Proprietary formats
- Many many formats
- Specific skill sets and expertise required
- Interoperability - Different tools render your content differently!
- More!

Different kinds of digital

- Digital tapes that do not present original encoding (Digital Betacam)
- Digital tapes that present the original encoding (DV/DAT/Betacam SX)
- File-based media
- Complex Packages (DCP/XDCAM/P2)
- Optical digital (some optical media is analog!) media
- Digitally re-formatted files from analog sources (VHS, UMATIC, Betacam SP)
- For more classes, see IASA TC-06 - https://www.iasa-web.org/tc06/guidelines-preservation-video-recordingss

Physical Formats

- http://www.arts.texas.gov/wp-content/uploads/2012/04/video.pdf is an excellent guide to figuring out what video formats you hold. Helps to communicate with a vendor and get accurate quotes.
- https://www.archives.gov/preservation/formats/motion-picture-film-identify-formats.html Library Of Congress simple guide on how to give a broad identification of film holdings.

Formats

- Question - What formats are you aware of?
- How have you encountered them?
- Include physical formats and digital file types.
- Be as specific as you like!

File Formats

- A single file format identification for the whole file is not sufficient
- AV is complex
- Multiple encodings present in a single file
- Some packages, like DCP/P2/XDCAM, defy traditional identification tools as they span multiple file types

PRONOM - Matroska with uncompressed video

```
kieranjol@kieranjol-ThinkPad-T420:~$ sf '/home/kieranjol/matroska_example.mkv'
---
siegfried   : 1.7.8
scandate    : 2018-05-27T18:11:42+01:00
signature   : default.sig
created     : 2017-12-02T14:49:15+11:00
identifiers :
  - name    : 'pronom'
    details : 'DROID_SignatureFile_V93.xml; container-signature-20171130.xml'
---
filename : '/home/kieranjol/matroska_example.mkv'
filesize : 27650108
modified : 2018-05-27T14:19:45+01:00
errors   :
matches  :
  - ns      : 'pronom'
    id      : 'fmt/569'
    format  : 'Matroska'
    version : '1-4'
    mime    :
    basis   : 'extension match mkv; byte match at 0, 41'
    warning :
```

PRONOM - Matroska with lossy, proprietary Apple ProRes video

```
kieranjol@kieranjol-ThinkPad-T420:~$ sf mpeg.mkv
---
siegfried   : 1.7.8
scandate    : 2018-05-27T18:14:26+01:00
signature   : default.sig
created     : 2017-12-02T14:49:15+11:00
identifiers :
  - name    : 'pronom'
    details : 'DROID_SignatureFile_V93.xml; container-signature-20171130.xml'
---
filename : 'mpeg.mkv'
filesize : 139566
modified : 2018-05-27T18:11:13+01:00
errors   :
matches  :
  - ns      : 'pronom'
    id      : 'fmt/569'
    format  : 'Matroska'
    version : '1-4'
    mime    :
    basis   : 'extension match mkv; byte match at 0, 41'
    warning :
```

```
General
Unique ID                                : 97249390257844370707038143158112476034 (0x49298D207EF42606A994DE5E4E53DB82)
Complete name                            : ffv1_pcm.mkv
Format                                   : Matroska
Format version                           : Version 4 / Version 2
File size                                : 13.0 MiB
Duration                                 : 2 s 0 ms
Overall bit rate                         : 54.6 Mb/s
Writing application                      : Lavf58.10.100
Writing library                          : Lavf58.10.100
ErrorDetectionType                       : Per level 1

Video
ID                                       : 1
Format                                   : FFV1
Format version                           : Version 3.4
Codec ID                                 : V_MS/VFW/FOURCC / FFV1
Duration                                 : 2 s 0 ms
Width                                    : 640 pixels
Height                                   : 480 pixels
Display aspect ratio                     : 4:3
Frame rate mode                          : Constant
Frame rate                               : 25.000 FPS
Color space                              : RGB
Bit depth                                : 8 bits
Scan type                                : Progressive
Compression mode                         : Lossless
Writing library                          : Lavc58.14.100 ffv1
Default                                  : Yes
Forced                                   : No
coder_type                               : Golomb Rice
MaxSlicesCount                           : 4
ErrorDetectionType                       : Per slice

Audio
ID                                       : 2
Format                                   : PCM
Format settings                          : Little / Signed
Codec ID                                 : A_PCM/INT/LIT
Duration                                 : 2 s 0 ms
Bit rate mode                            : Constant
Channel(s)                               : 1 channel
Sampling rate                            : 48.0 kHz
Bit depth                                : 24 bits
Writing library                          : Lavc58.14.100 pcm_s24le
Default                                  : Yes
Forced                                   : No
```

Core Trust Seal

- Confidentiality/ethics:
- Essential to know all the information contained in your files:
  - Sensitive information in metadata
  - Sensitive information in audio tracks

Containers and codecs

- AV objects are usually comprised of streams within containers
- Container examples: MOV/AVI/MXF/MPEG/MKV/MP4
- Many different kinds of streams can be stored in containers:
- Video streams: Uncompressed/H264/ProRes/Mpeg-2/FFV1/JPEG2000
- Audio streams: Uncompressed PCM/FLAC/MP3/Mpeg-4
- Some containers are flexible (Matroska)
- Some are not (MP4)
- Format analysis - DROID/PRONOM will only reveal container, not streams
- A codec may become a preservation risk before a container, and vice versa

What does a container do?

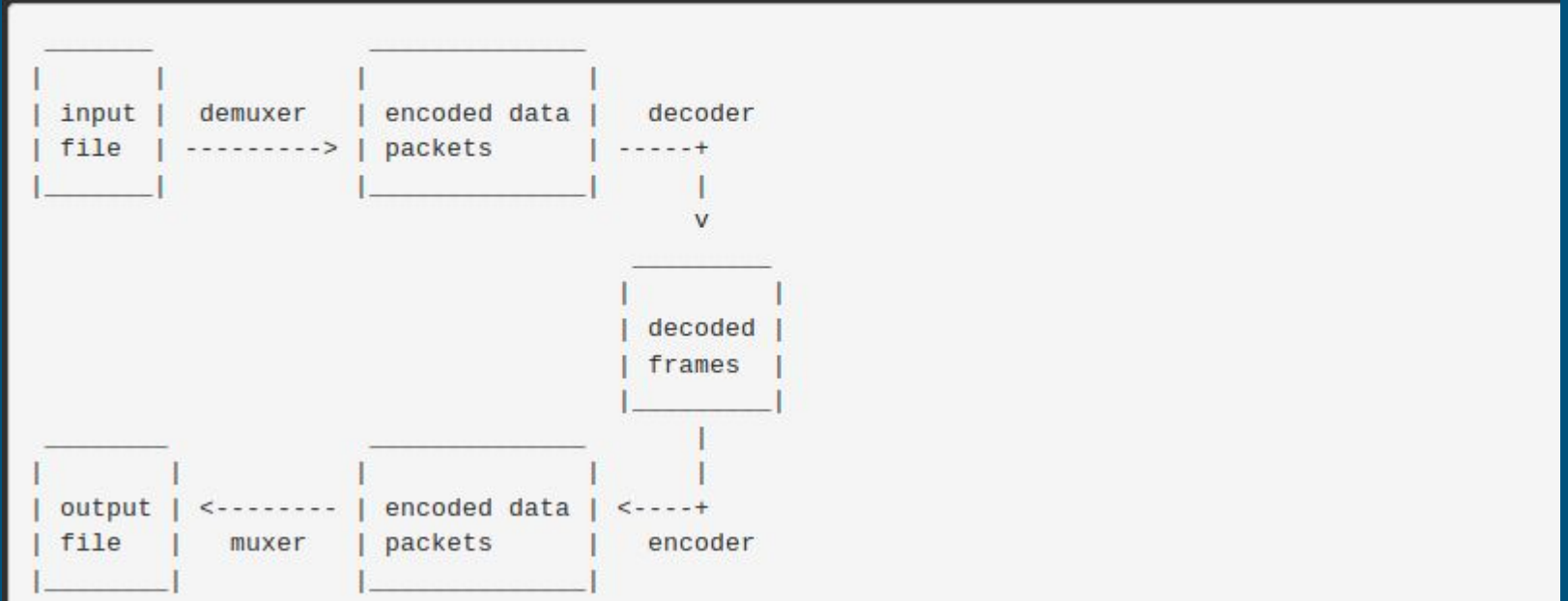
- Stores streams (audio/video/data/subtitles)
- Provides context and metadata
- Tells video players how to display AV correctly
- Containers sometimes contain the wrong metadata!
- Some codecs contain this essential metadata, most codecs do not!
- Codecs do not affect resolution or determine ‘quality’
    - BUT! - Containers can stipulate aspect ratio, interlacement, colour metadata that can *transform* the stored AV streams

What does a codec do?

- CoDec (enCOde/DECode)
- Stores (Encodes) and renders (decodes) the actual numerical AV data
- May require container to successfully render (some codecs don’t need a container!)
- May compress, may not
- Even uncompressed video can perform ‘efficiencies’ (packing 10-bit data)

3 Detailed description

The transcoding process in `ffmpeg` for each output can be described by the following diagram:



https://ffmpeg.org/ffmpeg.html

Question

- From what you've seen so far, how is AV ***similar*** to image/document formats that you might have encountered?

Mediainfo

- Open Source
- GUI/Command line/web application
- Information on technical specifications of your videos
- Render metadata in various levels of verbosity
- ESSENTIAL for quality controlling the work of vendors.
- Render metadata in:
  - MPEG-7 XML
  - PBCore XML
  - EBUCore XML
  - JSON
  - More!

Exercise

- You institution has specified the following technical characteristics - use mediainfo to see which passes and fails:
  - Video codec - FFV1
  - Container - Matroska
  - Interlacement - Top Field First
  - Frame Size - 720x576
  - Aspect Ratio - 4:3
  - Framerate - 25
  - Audio - 24-bit, 48kHz, Little Endian
  - Colour Primaries - BT.601
  - Bit Depth - 10
  - Chroma Subsampling 4:2:2

Note on Bit Depth

- Range of potential values from low to high
- 8-bit (consumer video) = 256 potential values from black to white
- 10-bit (Professional/archival video) = 1024 potential values from black to white
- Think of it like an artists palette:
  - Black and white look the same whether video is 1/8/10/16-bit
  - More intermediate, precise gradations possible with higher bit depth

Preservation goals

- Maintain the intended rendition of the significant properties of your video
- Maintain integrity over time
- Maintain accessibility
- Understand your objects
- Plan for potential migration

Digitisation Vendors

- What is desirable in a vendor?

Digitisation Vendors

- Set clear requirements
- Ensure that requirements are met
- Use VARRFP (http://memoriav.ch/wp-content/uploads/2014/07/VARRFP.pdf) as template for building a RFP
- Create quality control criteria
- Communication - question Quality Control issues
- Perform advocacy - the vendor may need guidance
- Ask for checksums
- Ask for the documentation of the digitisation environment
- Keep the ‘different classes of digital video’ in mind.

VARRFP - Sample RFP summary

1. Project Description - What is the collection, what formats?
2. Technical Requirements:
    a. File Formats
    b. Master/Mezzanine/Proxy
    c. Digitisation environments

Metadata

2.2 - Metadata:

Exercise - Take 5 minutes to write down the kinds of metadata that a digitisation might be expected to provide/what would you like to receive?

Metadata 2.2

- Confirm provenance metadata about source tapes provided by your institution
- Technical metadata of digitised files
- Environment information
  - Consider PREMIS events and agents vocabulary
  - Agents: Hardware/Software/People/Institutions
  - Events:
    - Creation/Migration/Normalisation
    - Metadata extraction
    - Message Digest Calculation
    - Fixity Check

VARRFP continued..

3. Vendor Information

- Ask vendor to comply with collections management requirements of your media. Specifications include:
  - Temperatures/humidity
  - Skilled engineers
  - Registration of items and confirmation of delivery
  - Maintenance of equipment

VARRFP - Workflow and transfer specs

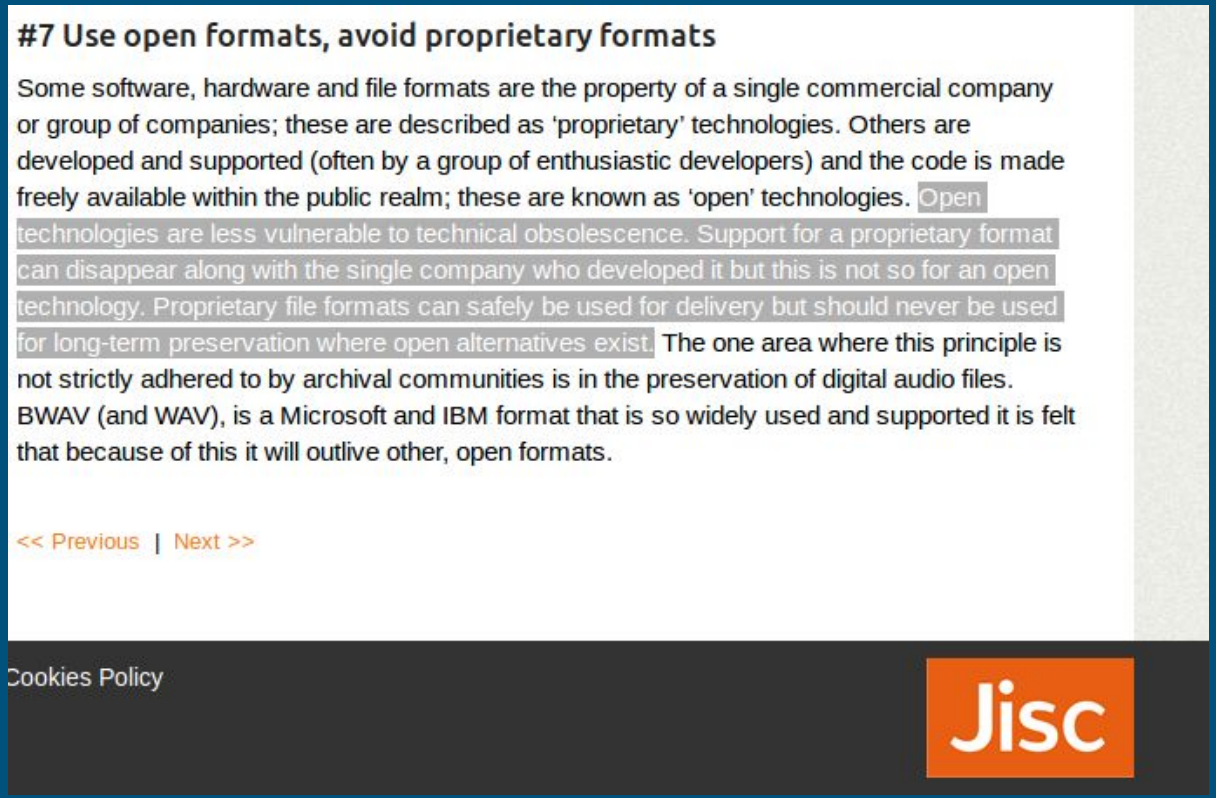
Section 3.1.2

- General goal: Accurately migrate the significant properties of the source material. Very technical section, but very important.

Section 3.1.2.1:

- Vendor QA - Checksums, validation, QA

JISC Digital Media



#7 Use open formats, avoid proprietary formats

Some software, hardware and file formats are the property of a single commercial company or group of companies; these are described as 'proprietary' technologies. Others are developed and supported (often by a group of enthusiastic developers) and the code is made freely available within the public realm; these are known as 'open' technologies. Open technologies are less vulnerable to technical obsolescence. Support for a proprietary format can disappear along with the single company who developed it but this is not so for an open technology. Proprietary file formats can safely be used for delivery but should never be used for long-term preservation where open alternatives exist. The one area where this principle is not strictly adhered to by archival communities is in the preservation of digital audio files. BWAV (and WAV), is a Microsoft and IBM format that is so widely used and supported it is felt that because of this it will outlive other, open formats.

<< Previous | Next >>

Jisc

Which file format?

- Ask the experts: AMIA/FIAF/Memoriav/NYU/UCLA and more! - Possibly useful for CTS VI. Expert Guidance.
- When in doubt:
- 10-bit Uncompressed video (v210) for analog video.
  - FFV1/Matroska or JPEG2000/MXF if you know what you're doing.
- QuickTime/MOV container
  - Again, Matroska - if you know what you're doing
- 24-bit, 48kHz little endian uncompressed PCM audio in WAV container
  - 96kHz, 24-bit little endian uncompressed PCM audio in WAV container for non video material.
- Think about validation - do you have valid, well formed files?

But not so fast!

- Significant properties - what kind of source material is it? CTS VII Data integrity and authenticity
- DV - retain DV encoding, ask for DV via firewire stored in MOV
- HDV - retain MPEG-2 encoding via firewire
- DAT - 16-bit audio (24-bit is usually the ‘archival standard’)
- If the tape deck does not allow you to capture original signal:
  - 10-bit Uncompressed Video or FFV1 lossless
  - 24-bit, 48kHz PCM audio
  - QuickTime or Matroska Container

XV. Technical infrastructure

R15. The repository functions on well-supported operating systems and other core infrastructural software and is using hardware and software technologies appropriate to the services it provides to its Designated Community.

Compliance Level:

Response

Guidance:
Repositories need to operate on reliable and stable core infrastructures that maximizes service availability. Furthermore, hardware and software used must be relevant and appropriate to the Designated Community and to the functions that a repository fulfils. Standards such as the OAIS reference model specify the functions of a repository in meeting user needs.

For this Requirement, responses should include evidence related to the following questions:

- What standards does the repository use for reference? Are these international and/or community standards (e.g., Spatial Data Infrastructure (SDI) standards, OGC, W3C, or ISO 19115)? How often are these reviewed?
- How are the standards implemented? Are there any significant deviations from the standard? If so, please explain.
- Does the repository have a plan for infrastructure development? If so, what is it?
- Is a software inventory maintained and is system documentation available?
- Is community-supported software in use? Please describe.
- For real-time to near real-time data streams, is the provision of around-the-clock connectivity to public and private networks at a bandwidth that is sufficient to meet the global and/or regional responsibilities of the repository?

Quality Control

- Quality Control the metadata with MediaInfo
- Quality control the AV content with QCTools
- Evaluate your media files
- Use AVAA (AV Artifact Atlas) to identify errors and develop a vocabulary: https://bavc.github.io/avaa/index.html
-

QCTools

- Developed by Bay Area Video Coalition and MediaArea/Dave Rice/Ashley Blewer
- Quality control migrated videotapes
- Graph/player view
- Great for communicating issues to vendors
- Can be more effective than visual QC, especially for non-specialists.
- Open source!
- Basic instructions: ***Look for big spikes in the following graphs, and investigate!:***
    - SAT, TOUT, MSEF, PSNR, VREP, Y/U/V, DIFFs

Migration/normalisation

- Normalise, or accept everything into your collection? (CTS VIII - Appraisal)
- How to verify authenticity of normalisation? Can you garner sufficient trust (CTS VII)
- How to garner trust with designated community?
- Framemd5/md5 and limitations https://amiaopensource.github.io/ffmprovisr
- Can your destination format handle the properties of your source file? - ffmpeg -h
- Is all metadata intact?
- *Is it worth the effort and can you actually stand by the results?*

Resources

- IASA TC-06: https://www.iasa-web.org/tc06/guidelines-preservation-video-recordings
- Videotape Identification: http://www.arts.texas.gov/wp-content/uploads/2012/04/video.pdf
- Basic Film Identification: https://www.archives.gov/preservation/formats/motion-picture-film-identify-formats.html
- RFP Guide and Template (VARRFP): http://memoriav.ch/wp-content/uploads/2014/07/VARRFP.pdf
- QCTools help: http://bavc.github.io/qctools/
-