The Perils of Preserving Digitized and Born-Digital Video

A Presentation for the Council of Intermountain Archivists

Fall, 2011

Jimi Jones
Digital Audiovisual Formats Specialist
Office of Strategic Initiatives
Library of Congress
jjones@loc.gov
http://www.digitizationguidelines.gov/

Analog vs. Digital

- Analog = signal is based in voltage
- Digital = stream of digital numbers decoded by computer

Digitized vs. Born Digital

- Digitized = analog source information converted to digital information
- Born Digital = information was initially captured as digital and stays that way

3 Basic Steps to Digitization

- Sampling
- Quantizing
- Coding

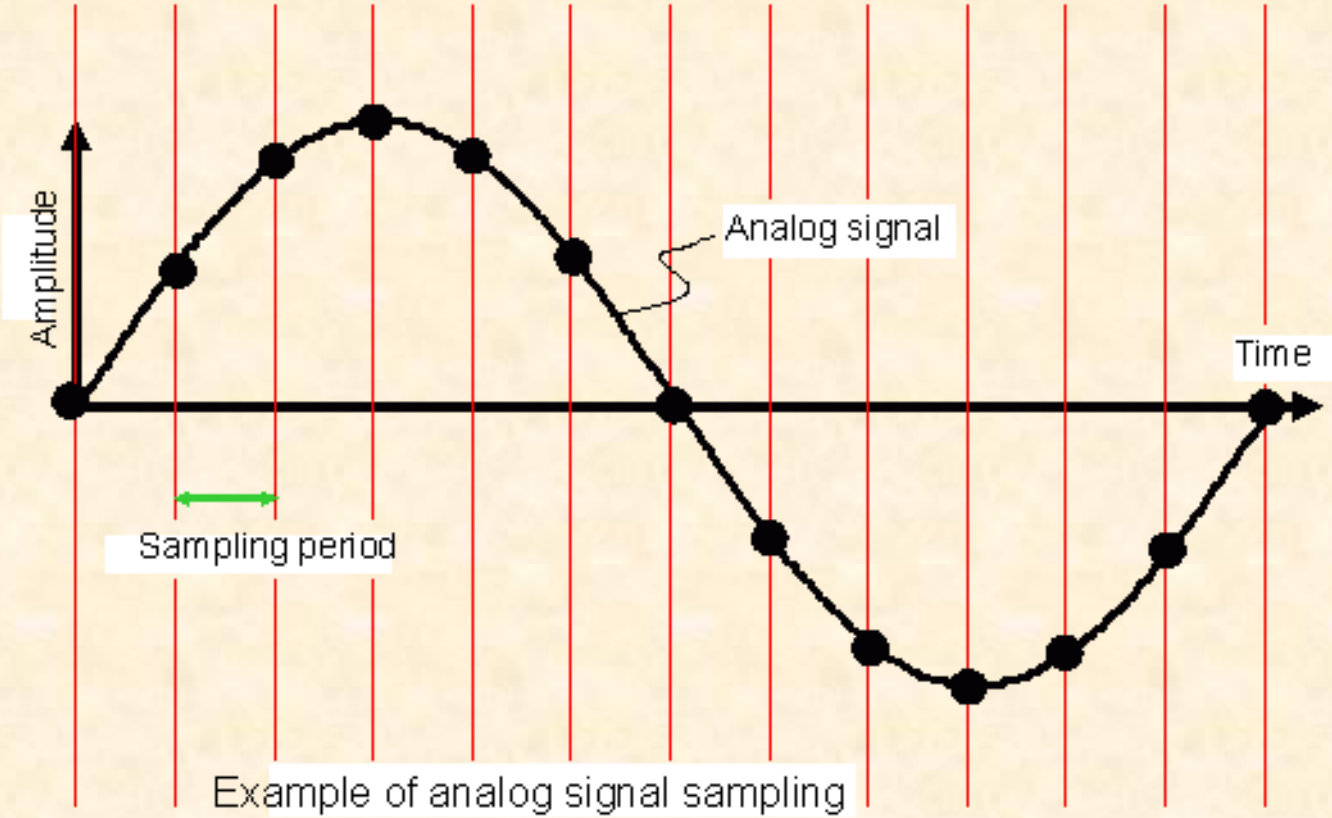
Sampling

- A sample is essentially the computer "checking in" to see where the audio/video signal is
- The more you sample, the better the "resolution" or quality of the recording
- Analog waveform is measured at regular intervals – each interval of measure is a sample

Sampling

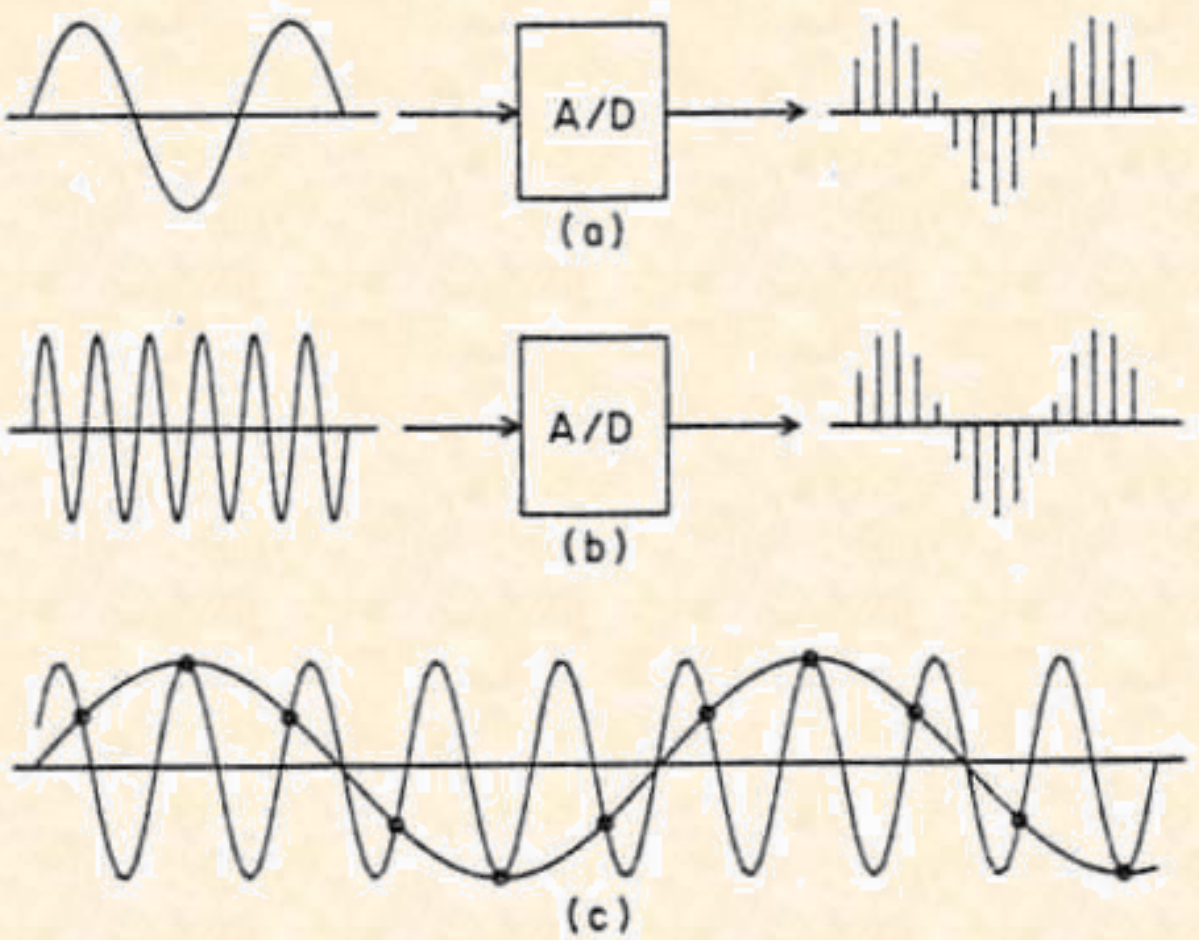
- The more you sample, the more information there is in the file
- The more information in the file, the larger the file
- The more “space” between samples, the more the computer is guessing at the wave’s form

Sampling



Example of analog signal sampling

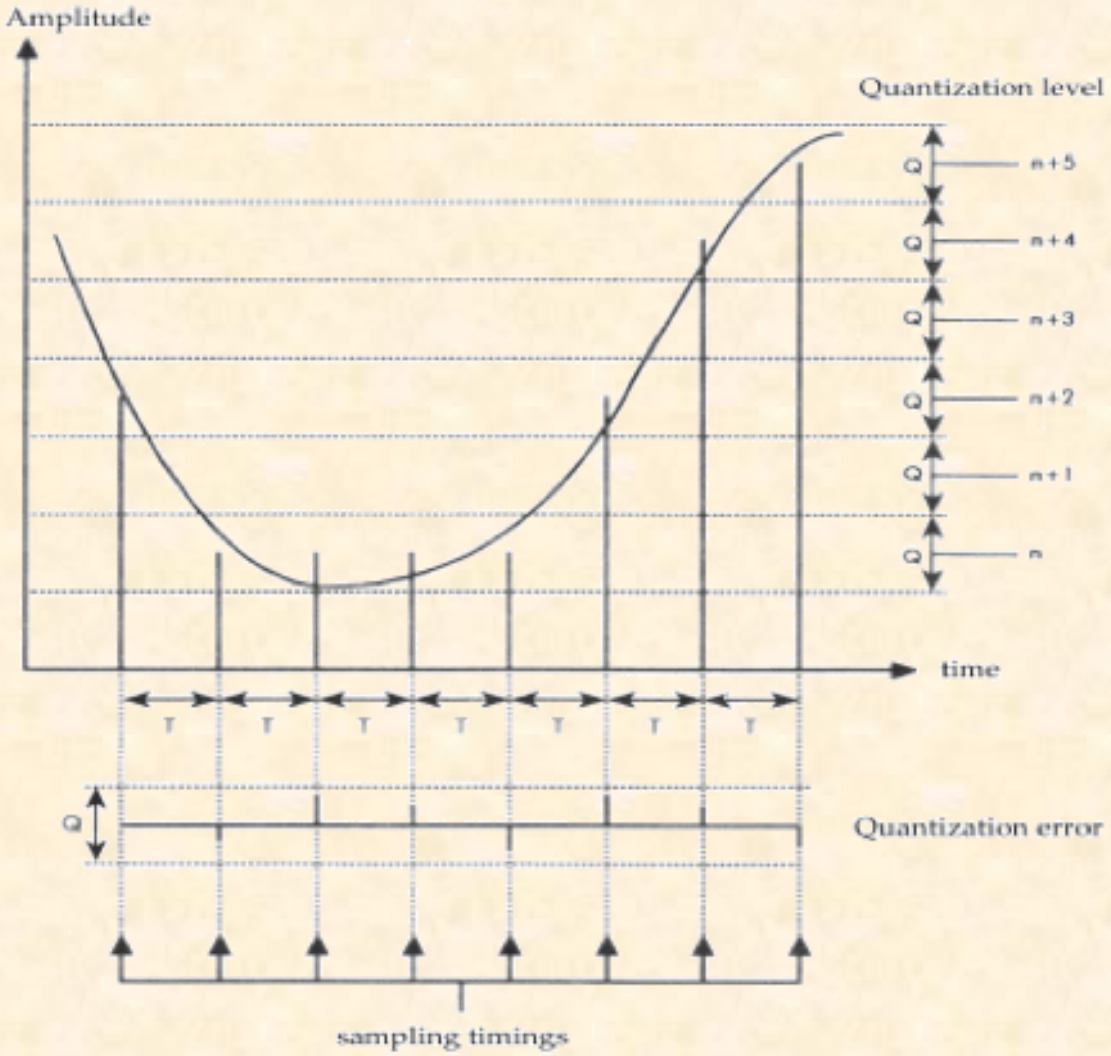
Sampling



Quantizing

- For each sample, decisions are made by the analog-digital converter as to what digital value will be assigned to the sample.
- For example, if a voltage falls between two numbers available to represent the voltage digitally, a decision will be made as to whether to go with the higher or lower number.

Quantizing



Coding

- Assigning numerical value to the sample/signal
- This is the information that makes up the bulk of the AV bitstream

Digital Video

- Luma – brightness (one luma component)
- Chroma – color (two chroma component)
- Frame Rate – number of image frames displayed per second
  - Lower frame rate, smaller file but poorer quality image
- Bitrate (or bit rate) – is a measure of the rate at which the information is moving in a video stream
  - The faster the bit rate, the better the image quality
  - Can be variable or fixed
    - Fixed moves at a constant rate
    - Variable – rate changes based on the complexity of image (the more complex the image, the slower it goes)

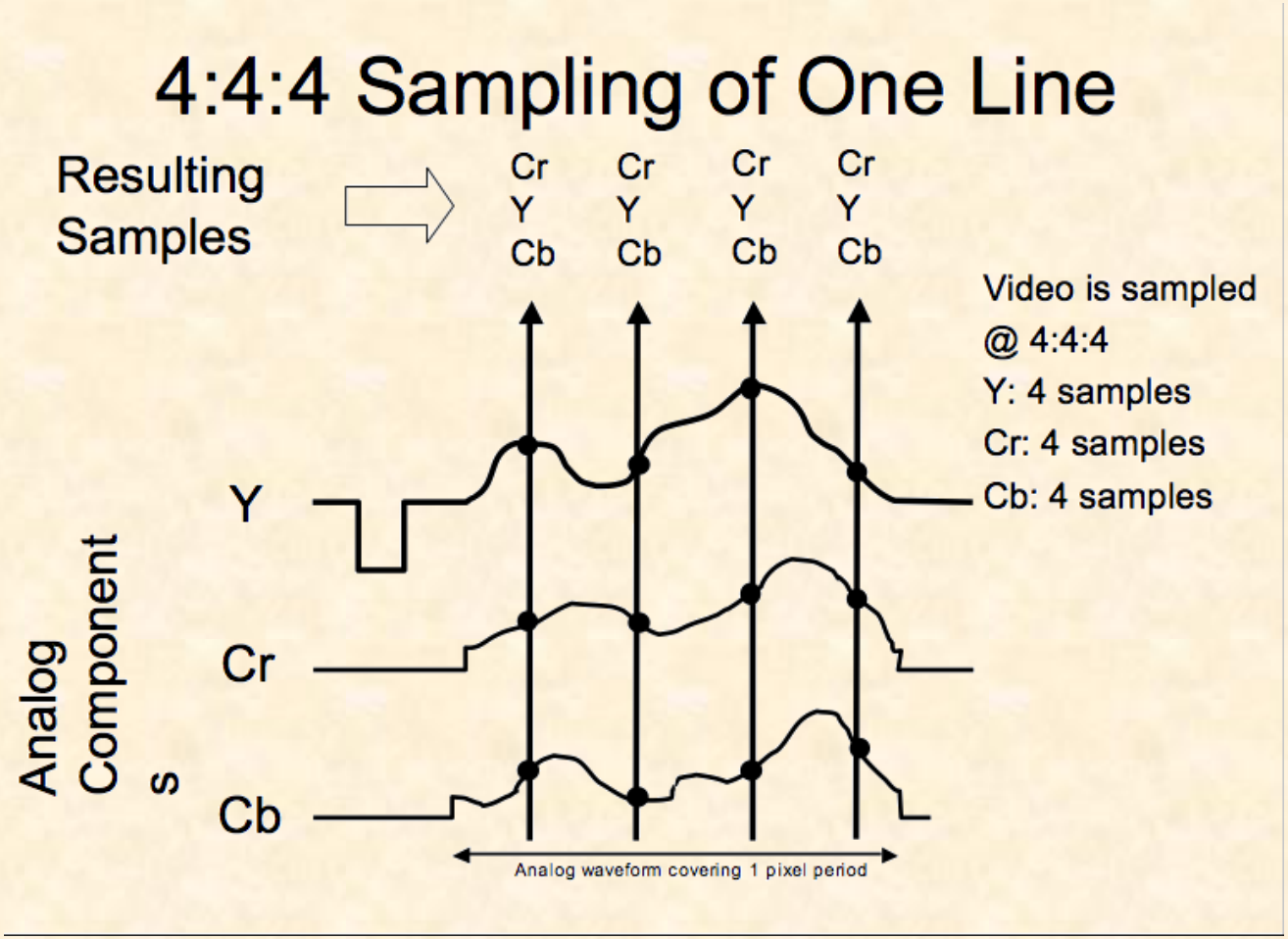
Video Sampling

- Expressed as a relation of luma to red chroma to blue chroma information (ex: 4:2:2)
- Each number represents how many times each component is sampled per fixed unit of time

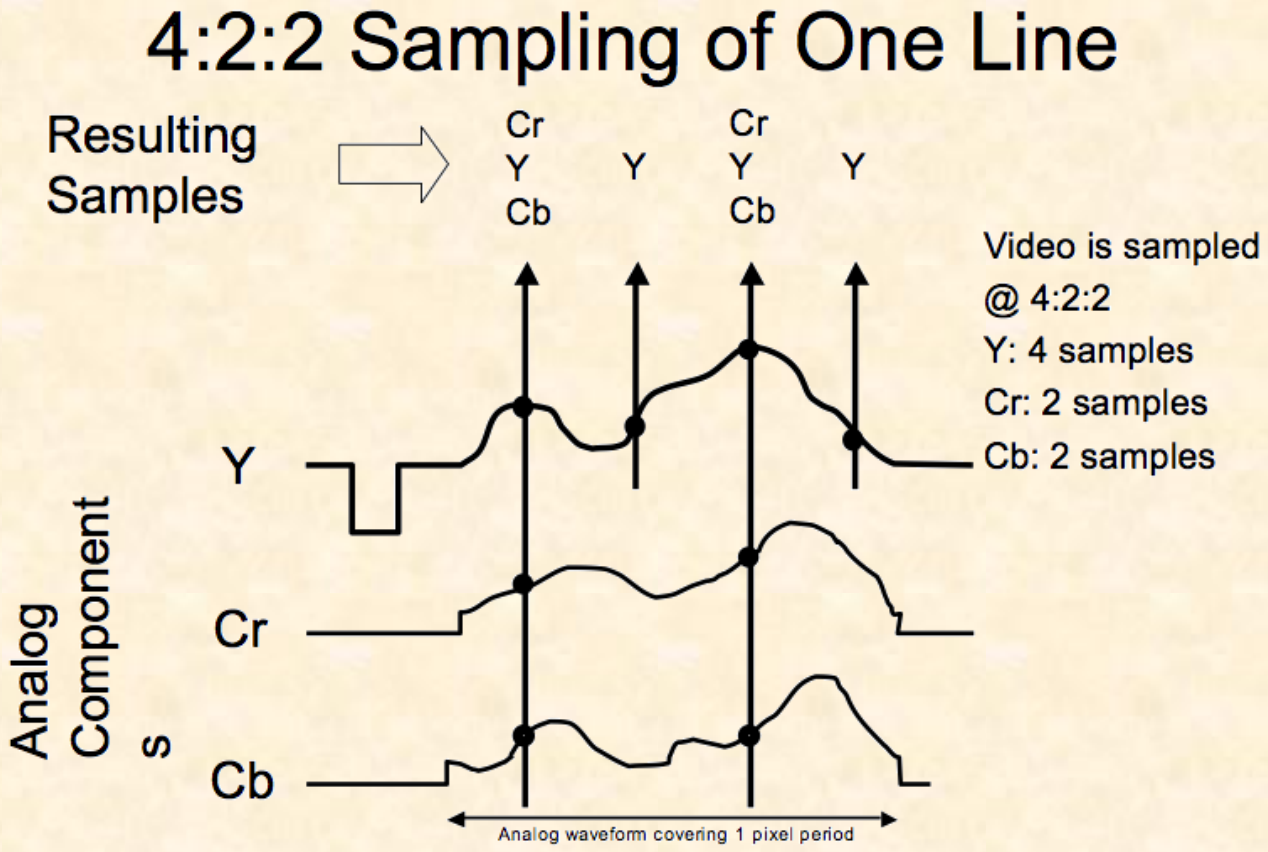
Video Sampling

- 4:2:2 (a common broadcast “color space”) means the luma is sampled 4 times and each chroma is sampled twice
- 4:2:2 also means that the luma (brightness) is sampled twice as much as the color information
- 4:4:4 would be the highest sampling rate; the largest kind of file; the best image quality

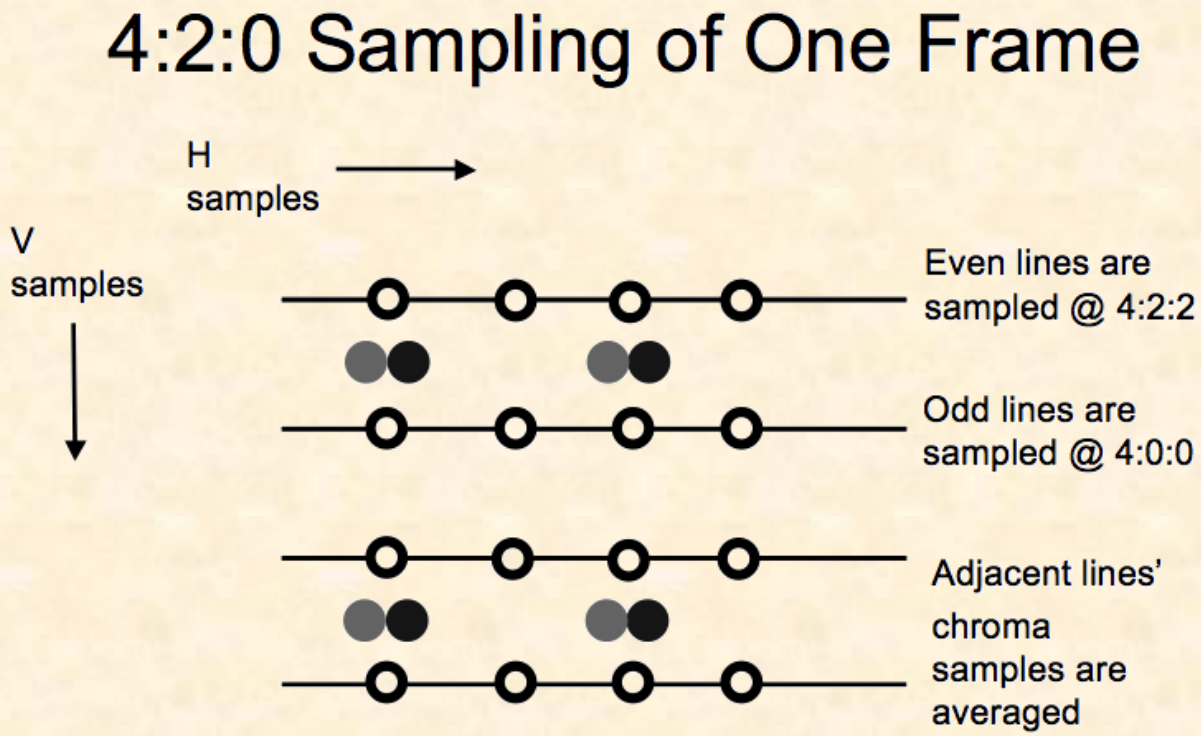
4:4:4 Sampling



4:2:2 Sampling (Broadcast Space)



4:2:0 Sampling (Many consumer digital video formats)



What is compression?

- Compression makes files smaller and easier to use
- In video, compression discards color information, resolution and, sometimes sound quality
- Once a file is compressed, that information cannot be recovered
- Compressed files are not considered archival quality

What is compression?

- Once a file is compressed, that information cannot be recovered (except J2K)
- Compressed files are not considered archival quality
- Lossless compression – file size is diminished without loss of information
- Lossy compression (most common) – file size is diminished by discarding information

Video Compression Types

- Int**er**frame
  - Across a group of frames (GOP)
  - Keyframe, then other frames are compared to that – change difference is recorded, not discrete frames all the way through the video
  - MPEG video uses interframe compression
  - Can’t edit this video with transcoding

Video Compression Types

- Int**ra**frame
  - Each frame is compressed discretely within itself (image compression)
  - DV (very common consumer/prosumer format) uses intraframe compression so it can be edited easily – cut on frames
  - Uses DCT (Discrete Cosine Transform) to compress video frame-by-frame

Video Compression

Codecs

- Codec: Coder/Decoder – a computer program that create and/or play back digital video and/or audio
- Also can stand for "Compressor/Decompressor"
- Unlike the audio world, video has no archival standard codec yet (JPEG200 is promising)

Codecs

- Creates the “bitstream” or “essence” of the file
- http://en.wikipedia.org/wiki/Category:Video_codecs
- http://en.wikipedia.org/wiki/Category:Audio_codecs

Some Common Video Codecs

- WMV (Windows Media Viewer) – proprietary
- DivX – proprietary
- Sorenson (used in Quicktime) – proprietary
- Xvid – open source
- Real Video - proprietary

Video Wrappers

- Contain the video essence information (the bitstream)
- Wrap the bitstream up with information that describes it and explains how to decode it
- Sometimes we say “file format” in place of “wrapper” – not entirely accurate but common

Common Video File Wrappers

- Quicktime MOV – proprietary; common on Apple platforms; well supported
- AVI (Audio Video Interleaved) – proprietary; common on Windows platforms; well supported
- Matroska – open source; free; support is growing
- MXF (Material Exchange Format) – defined by SMPTE, used in the broadcast industry and at LOC; support is growing

MPEG

- MPEG-2 – standard devised by Motion Picture Experts Group
  - DVD encoding
  - Compressed
  - Used in broadcast
- MPEG-4 – used online as downloadable video
  - very compressed
  - Used in iPod/iPhone/iPad video

JPEG 2000

- Lossless video compression – discards no information (original data can be reconstructed from compressed data)
- Large file size becomes smaller when encoded as J2K
- Adoption is spotty but growing
- Has space for metadata
- Averages about 2.3:1 compression, depending on how “busy” the image content is

Digital Preservation Parallels Analog AV Preservation

- Must preserve playback software alongside media files
- Many institutions reformat their digital files
- Masters and access copies
  - MXF-wrapped J2K – master
  - MOV H.264 15 Mbps – access copy

What the *&%$ Do We Do With Born Digital?

- File formats (wrappers) are rapidly evolving
- How do we decide which formats are potentially long-lived and which are at risk?

What the *&%$ Do We Do With Born Digital?

- Keep a copy of the file in the format in which it was received (Internet Archive does this)
- Transcode (change the file type) as necessary
- Transcoding must be a very informed decision – can cause unnecessary “bloat” to files by inserting empty data (bigger files don’t necessarily mean higher quality)

What the *&%$ Do We Do With Born Digital?

- LOC has engaged George Blood and Safe Sounds Archive to help assess which format types can be kept in situ and which need to be transcoded
- Can have to do with potential obsolescence as well as “fragility” of file structure (DVD’s VOB for example)

George Blood Interim Digital Video Format Report for FADGI

Category 2. Digital source (media dependent, non-transcoded transfer possible)

Recommended delivery specification

Method: Transfer data to file wrapper without transcoding
Wrapper: Native-as-associated with the underlying encoded essence, (such as .dv, .imx, .mpeg, .mp4, etc.), or .mov (QuickTime) or .avi[1]. Otherwise wrap in QuickTime.

Video:

- Video Compression: Native
- Frame size: Standard definition: 525: 720x486[2]
  625: 720x576
  High Definition: Native
- Aspect Ratio: Native: Standard Def: 4:3
  High Def: Native
- Bit Depth: Native, 8-bit or 10-bit
- Color Space: 525: YCbCr
  625: YUV
- Chroma subsampling: Native, typically, 4:2:2, 4:1:1, or 4:2:0
- Interlaced/Progressive: same as original format
- Frame Rate: Native: Standard Def: 525: 29.97
  625: 25
  High Definitnion: Native

Time Code: Native when present; Midnight start & NDF if synthetic

Audio:

- Audio channels: Same as original
- Audio Compression: Native, typically uncompressed, PCM
- Audio Sample Rate: Native, typically, 48 kHz
- Audio Resolution: Native, typically 24- or 16-bit

George Blood Interim Digital Video Format Report for FADGI

Category 5. Digital source ("authored" DVDs and other "authored" disks)

Recommended delivery specification (authored DVDs)

Method: extract VOB files from disk
Wrapper: retain VOB
Video:
- Video Compression: Native, MPEG2
- Frame size: 720 x 480, rectangular pixels
- Resolution: 8-bit
- Color Space: Native, YCbCr
- Chroma subsampling: Native, 4:2:2
- Interleaved/Progressive: same as original format
- Frame Rate: 29.97 NDF, or same as original (no-inverse telecine)[1]

Time Code: Midnight start
Audio:
- Audio channels: same as original
- Audio Compression: Native, typically Dolby Digital
- Audio Sample Rate: Native, typically, 48 kHz
- Audio Resolution: Native, typically 16 bit

Obsolescence Monitoring Comment: Moderately stable, one transcode may be necessary in 3-10 years before JPEG2000/MXF migration. Monitor carefully since VOB may become obsolescent in the near- to medium term.

Digital Media Preservation

- Must be part of an ongoing preservation workflow
- There is no **end** to digital preservation (or any preservation)
- Checksums can look at the data in a file and see if there is something missing or corrupted

Migration

- Move file information from older version of a file format to a new one
- Typically involves moving a large number of old files to a smaller number of contemporary formats
- Usually about a 5-10 year migration schedule

What Kinds of Wrappers/Formats Do We Want?

- Open standards
- Widely supported
- Doesn’t have patent or other kinds of restrictions to use
- Ideally, something that is easily transcoded
- Has minimal external dependencies – on particular software and/or hardware

Thoughts About Digital Materials and Preservation

- Consider your digital storage solution
- RAID drives back up information and prevent total data loss
- Many larger collections are using tape-based long-term storage
- Must preserve playback equipment to digitize media in-house

A RAID Array


"R0" DRIVE
Serial # D1000711102
"R1" DRIVE
Serial # D1000711102
"R2" DRIVE
Serial # D1000711102
"R3" DRIVE
Serial # D1000711102

Thoughts About Digital Materials and Preservation

- Physical storage media for digital files must still be preserved (disks, hardware, tapes)
- Consider the LTO tape – a commonly-used storage medium that is based in magnetic tape

Tape-Based Storage

- Magnetic tape that will be heir to the problems of magnetic tape
- HOWEVER – obsolescence progresses so quickly that it's more likely that tapes will need to be migrated than that the stewards will have to deal with sticky shed, etc.

Tape-Based Storage

- For long-term storage (stores the information linearly)
- Not good for continually-accessed, high-use materials (spinning discs are better)
- LC uses T-10000 tape, not LTO because the error rate has been shown to be less for T-10000

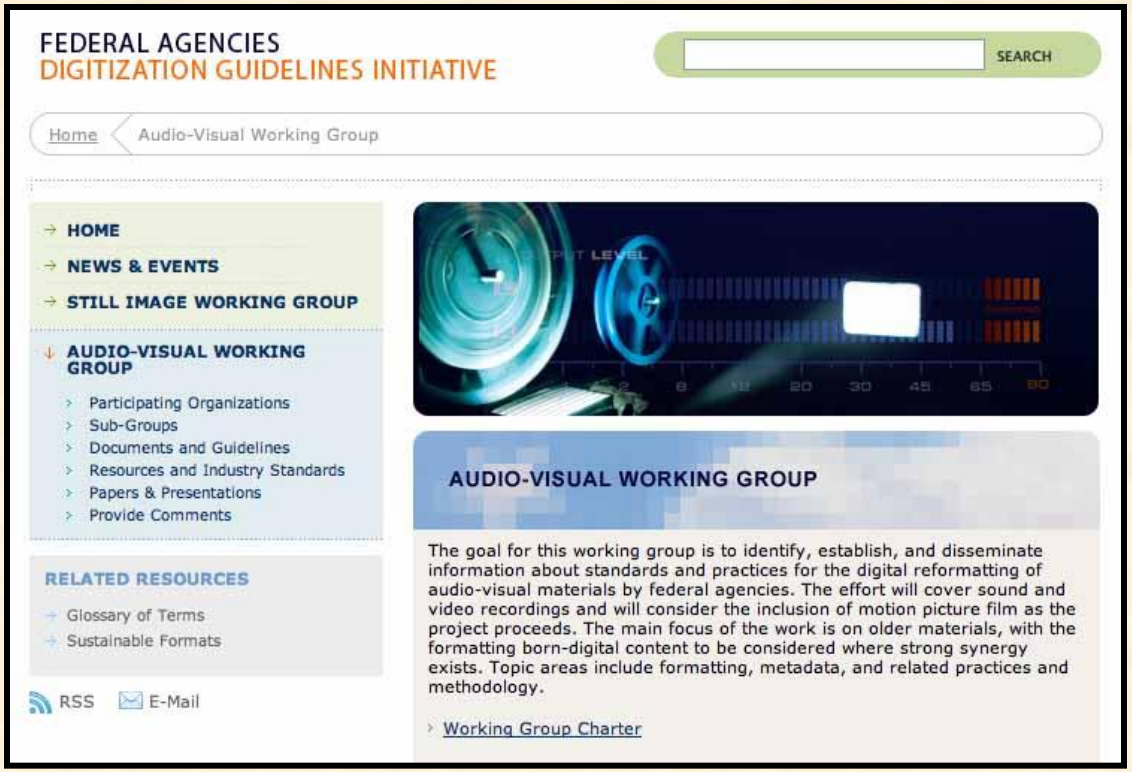
LTO Tape

The Future of Digital Video Stewardship?

- preserving our analog originals long enough to digitize them
- the “fire hose” of born-digital materials isn’t slowing – it’s increasing
- do we compete with YouTube, Vimeo, etc or collaborate with them?

The Packard Campus for Audio-Visual Conservation

SAMMA Video Digitization Machines

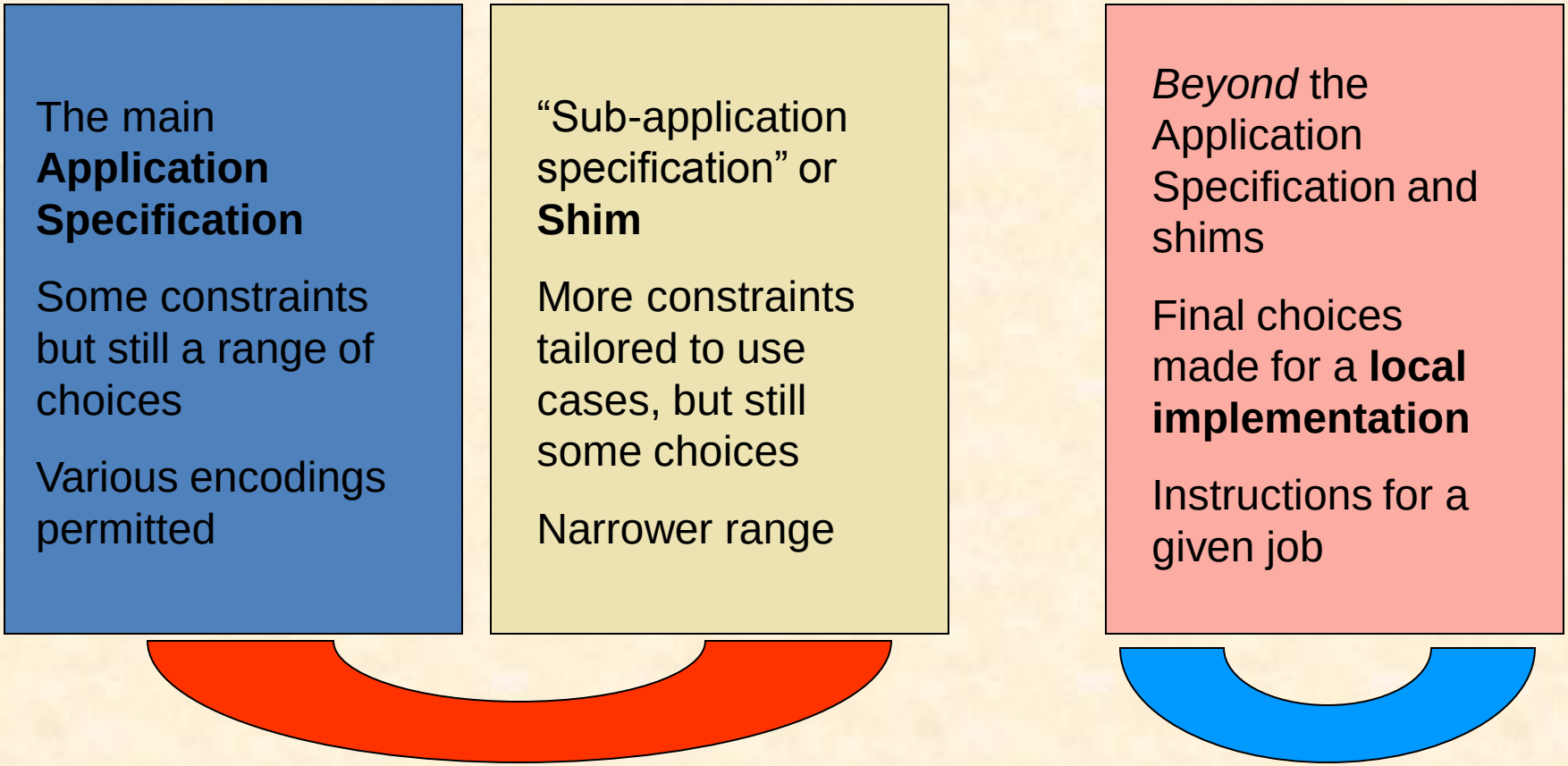


The Federal Agencies Audio-Visual Working Group

Memory institutions like LC and NARA are especially active

Participating agencies

MXF Application Specification, Shim, Local Implementation



Zone for the specification under development

Audio analogy: "BWF as constrained form of WAVE"

Zone for added advice, e.g., Fed Agency or IASA guideline

Audio analogy: "use 96/24 in your BWF"

Federal Agencies Digitization Guidelines Initiative (FADGI) MXF Application Specification

7.1 Picture

Different variants of this component may be selected by different Shims.

7.1.1 Picture – Compressed at Ingest (i.e, compressed by the archiving organization)

This parameter is typically selected by an archive that prefers to store a reduced-data file, and that is formatting or reformatting content as a part of its own pre-ingest or ingest activity, e.g., transferring content from a videotape carrier, or scanning film.

Dimension	Description	AS-AP Constraint	AS-AP Values
Program bitrate	how many bits per second at real time	Gentle	Up to 200 Mbps
Picture format	Picture raster and aspect ratio	Moderate	480i 4:3, 576i 4:3, 576i 16:9, 720p 16:9, 1080i 16:9, 1080p 16:9 2K 4K
Picture Essence Schemes	what picture signal schemes (compression or sampling or other) are encountered in programs	Gentle	JPEG 2000 broadcast profile Level 5 through level 7 [Other JPEG 2000 profile tbd] X'Y'Z' 10 – 16 bpp

7.1.2 Picture – Uncompressed (i.e., when produced by the archiving organization)

Federal Agencies Digitization Guidelines Initiative (FADGI) MXF Application Specification

7.1.4 Picture – Ancillary Still

This parameter is intended to permit the inclusion of image-based corollary materials associated with content that an archive is reformatting, e.g., documents or pictorial items stored with a source videotape. [The refined development of this specification will include other sections that address ancillary materials other than those that can be imaged.]

Dimension	Description	AS-AP Constraint	AS-AP Values
Program bitrate	how many bits per second at real time	Gentle	N/A
Picture format	Picture raster and aspect ratio	Moderate	Per input format
Picture Essence Schemes	what picture signal schemes (compression or sampling or other) are encountered in programs	Gentle	TIFF Others TBD

7.2 Sound

Dimension	Description	AS-AP Constraint	AS-AP Values
Sound Essence Schemes	what sound signal schemes	Moderate	PCM 96 kHz 24 bit
Sound Language repertoire	what primary sound languages may be present	None	N/A
Track Listings	what combinations of picture sound and data tracks are encountered in programs	Strong	Main Sound (1,2 or 6 channels) SAP (0 1 or 2 ch) DVS (0 1 or 2 ch)

Useful Links

- **The Library of Congress' Packard Campus for Audio-Visual Conservation: http://www.loc.gov/avconservation/packard/**

- **The Library of Congress' Sustainability of Digital Formats: http://www.digitalpreservation.gov/formats/**

- **The Federal Agencies Digitization Guidelines Initiative (FADGI) Audiovisual Working Group: http://www.digitizationguidelines.gov/audio-visual/**

- **More About the FADGI MXF Application Specification: http://www.digitizationguidelines.gov/guidelines/MXF_app_spec.html**