JPEG2000 in Moving Image Archiving

The Library of Congress – Packard Campus
National Audio Visual Conservation Center
Culpeper, Virginia

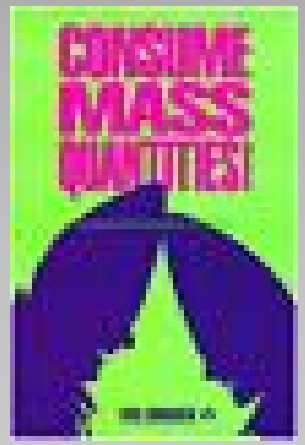
James Snyder
Senior Systems Administrator, NAVCC

Our Mission:
Digitize the Entire Collection

- Bring the ‘dark archive’ into the ‘light’
- Make it more accessible to researchers and the public
  - Stream audio and moving image to the public
  - Without compromising the rights of the © holders
- Required the development of new technologies & processes
- Required new ways of thinking about preservation

Digitizing Everything

Consume Mass Quantities!



Overall Digitizaton Goals

- Archive is essentially a permanent data set
  - We have a different perspective on the word 'longevity'
- Files should be able to stand alone without references to external databases
  - Lots of metadata wrapped in the files!
- Digitize items in their original quality
  - Audio at highest bandwidth in the recording
  - Video at its native resolution
  - Film scanned at resolution equal to the highest available on film
    - 4096 x 3112 for 16mm; 8192 x 6224 for 35mm in development

Overall Digitizaton Goals

- Use as much off-the-shelf products as feasible
- Invent items or write custom software only when a commercial product cannot fit the need
- Use industry standards as much as possible
  - Most video formats were created using standards anyway; why reinvent the wheel for preservation?

Overall Digitizaton Goals

- Video:
  - JPEG2000 'lossless' (reversible 5x3) (ISO 15444)
  - MXF OP1a
- Film: (planned)
  - JPEG2000 lossless with MXF OP1a wrapper (goal)
  - Short-term expediency: DPX with BWF audio
  - 4k (4096 x 3112) JPEG2000 Lossless encoding now available from vendors
  - Working with vendors to extend to 8k (and beyond, if needed)
- Audio:
  - 96 kHz/24 bit BWF RF64 (Broadcast Wave Format)
  - Limited metadata capabilities: looking at MXF OP1a wrapper for more metadata in audio-only files as well

Why JPEG2000 Lossless?

The first practical moving image compression standard that doesn't throw away picture content in any way.

Why JPEG2000?

JPEG2000

- An international standard (ISO 15444)
- The first, and currently only, standardized compression scheme that has a truly mathematically lossless mode
- No LA (Licensing Agreement) that has issues

Other compression systems

- No mathematically lossless profiles in any other standardized compression scheme
- Other have licensing agreements that have legal or temporal issues
- Many are vendor specific and thus used at the whims of the manufacturer

Why JPEG2000?

JPEG2000

- Unlike MPEG, it's a standard not a toolkit
- Wavelet based
- Can be wrapped in a standardized file wrapper (MXF) which promotes interoperability

Other compression systems

- MPEG is a toolkit, meaning different implementations exist for the same profile/level & bitrate
- MPEG, DV are DCT based, which results in unnatural boundaries that are very visible to the eye

Why JPEG2000?

JPEG2000

- Lossless means no concatenation artifacts created by encoding that aren't already there

Other compression systems

- All lossy compression schemes suffer concatenation artifacts
- Particularly bad between codecs that compress using different toolkits:
  - MPEG>DV
  - DV>MPEG
  - MPEG-4/H.264 > low bitrate MPEG-2

Why JPEG2000?

JPEG2000

- Can accommodate multiple color spaces
  - YPbPr
  - RGB
  - XYZ
  - New ones being worked on
- Can accommodate multiple bit depths
  - Video: 8 & 10 bits/channel
  - Film: 10-16 bits/channel

Other compression systems

- MPEG-2 is 8 bit ONLY
- MPEG-4: mostly 8 bit but one 10 bit profile for production
- YPbPr color space ONLY

File Format

MXF File Wrapper

Why Not JPEG2000 File?

- Part 3 defines a .jp2 moving image file, but it can't handle the variety of sources required in archiving
- MXF file standard already in progress, with far more flexibility designed into the standard

Solution:

Wrap JPEG2000 part 3 encoded moving image essence into the MXF file format

Why MXF?

Issue: Interoperability

- How to create files that will work across multiple platforms and vendors seamlessly?
- Most common production file formats today are both vendor specific:
  - .mov = Apple
  - .avi = Microsoft (original Windows video format)
- If the owner of the format decides to make a change or orphan the format: what then?

Interoperability Solution

- File format standardized by the SMPTE (Society of Motion Picture & Television Engineers) & AMWA (Advanced Media Workflow Association)
- Allows different flavors of files to be created for specific production environments
- Can act as a wrapper for metadata & other types of associated data

Interoperability Solution

- MXF: major categories are called “operational patterns” (OP)
- More focused subcategories are called “Application Specifications” (AS)
- Our version: OP1a AS-02
- Working on an archive-focused AS called AS-07 (aka AS-AP for Archiving and Preservation)

How We Implemented

SAMMA

SAMMA

- The Library was a driving force in the creation of the first production model JPEG2000 lossless video encoder: the SAMMA Solo
  - Can only do 525i29.97 and 625i25
  - Produces proxy files, but only in post-encoding process
- 31 currently deployed at Culpeper
- SAMMA Sync 'TBC' not really a TBC (time base corrector): it's a frame sync
  - Can't correct the worst videotape problems
  - Sometimes (rarely) injects artifacts into the video!
  - We use the Leitch DPS-575 TBC to correct our analog video problems. Corrects virtually any tape that can be read.

SAMMA

- The updated HD model premiered at NAB this past April.
  - Will encode both SD and HD and multiple frame rates including film frame rates
  - New SAMMA Sync still not a TBC, but MUCH better than the first version

Vendor Diversity

- Omneon & Amberfin have teamed up to create a video server based solution where multiple encoders feed one server
  - Can handle SD, HD & 2k at multiple frame rates
  - We will be using this solution for Congressional video
- OpenCubeHD currently shipping an encoding, editing and file creation platform
- DVS premiered 4k (up to 4096 x 3112) JPEG2000 Lossless editing and encoding at NAB in April
  - Including 3D @ 4k

Feature Diversity

- The entire production & distribution pieces are now in place:
  - Editing
  - Encoding & file creation
  - Metadata creation, editing & insertion
  - Proxy file creation at the same time
- Everything up to 3D

Future Needs

- Real time 4k encoding and decoding
- Encoding beyond 4k
  - 2011 NAB vendors had UHDTV and 8k film scanning as proposed or shipping products on the show floor
- Encoding of the new color spaces being proposed
- Finalize metadata needs: creation, editing & insertion toolkits in MXF files

We’re Not the Only Ones

Digital Cinema standardized on MXF for the distribution of Digital Cinema Packages to theatres

QC?

The goal is to QC every file we produce

Automated Software

- Interra Baton has a mature JPEG2000 Lossless automated QC package
  - Real time still a challenge; depends on computational power
  - We will be implementing this solution this year
- Tektronix Cerify is not quite as good as Baton, but getting better
- Digimetrix premiered their package at April's NAB and it shows promise

Error Detection

How do we know the files are good throughout the system, or later on?

SHA-1 Checksum

- Cryptographic Hash Checksums are designed to identify one bit flip in an entire file
- SHA-1 can accurately identify 1 bit flip in file sizes up to 2^{61} -1 bits
- First year of production: 800 TiB in the archive: no bit flips

Where Do We Store All This Material?

Our Digital Repository

- 200 TiB SAN
  - Staging area for transmission to backup site and the tape library
  - Backup site has identical SAN & tape library
- Tape library
  - StorageTek SL-8500 robot with 9800 slots currently installed; 37,500 total slots planned by ~2015 (expansion depends on requirements)
  - Currently using T10000-B tapes with 1TiB/tape current capacity (9.8PiB available; 37.5PiB total as designed)
  - Moving to new T10000-C tapes @ 5TiB/tape (eventually 49PiB available; 187.5 PiB total as designed)
  - Upgrade path to ~48TiB/tape by ~2019

The Digital Repository

- SAM-FS file system
- 1.35 PiB on tape as of Monday (5/9/2011)
- Increasing at approx. 20-40 TiB/week
  - 80-100 TiB/month
  - 60% of each month's production is JPEG2000 MXF files by data throughput
  - 20% of month's output is JPEG2000 MXF by file count
  - Total of 29,400 JPEG2000 files as of 5/9/2011
- First ExiB anticipated around or after 2020

Why T10000 tape?

Bit error rate matters!

Digital Repository Requirements

- Data is effectively a permanent data set
  - This is America's cultural archive
- Archive contents must stand on their own (no external databases required to know all about a file)
- Must be file format agnostic
- Must be scalable to very large size (EiB+)

Bit Error Rate Matters!

- When you get to the PiB level:
- 10^{-17} bit error rates is GiB of errors in your repository!
- T10k has best current error rate: 10^{-19}
- All other storage: currently the best is 10^{-17}
  - 2 orders of magnitude worse error rate!
- When you are migrating every 5-10 years your entire library, BIT ERROR RATE MATTERS!!!!

Issues

- Most commercial IT equipment has bit error rates of 10^{-14}, including Ethernet backbone equipment: what good is storage BER of 10^{-17} when your system's best BER is 10^{-14}
- How often to check data integrity?
  - Continuous above a certain size
  - Reading the data can also damage it!
- How often to migrate?
  - Individual files: every 5-10 years (we think)
  - Subject to verification!

Future Challenges?

Future Challenges

- New production systems coming online:
  - Congressional video archiving: 2-5 PiB/year?
    - 3300 hours x 720p59.97 HD/year
  - Born Digital file submissions: 2-5 PiB/year?
- HD video encoding now possible
- Live Capture system coming online

Future Challenges

- Standards work continues on…
  - SMPTE AXF proposed standard for media-agnostic file definition
  - SMPTE/AMWA MXF Application Specification for media files with extra metadata & associated essences enabled
  - Update the MXF standard to properly define JPEG2000 interlace vs. progressive video cadence
  - Work with AES, SMPTE, AMWA, AMPAS & others on defining a complete set of metadata standards (or at least templates!)

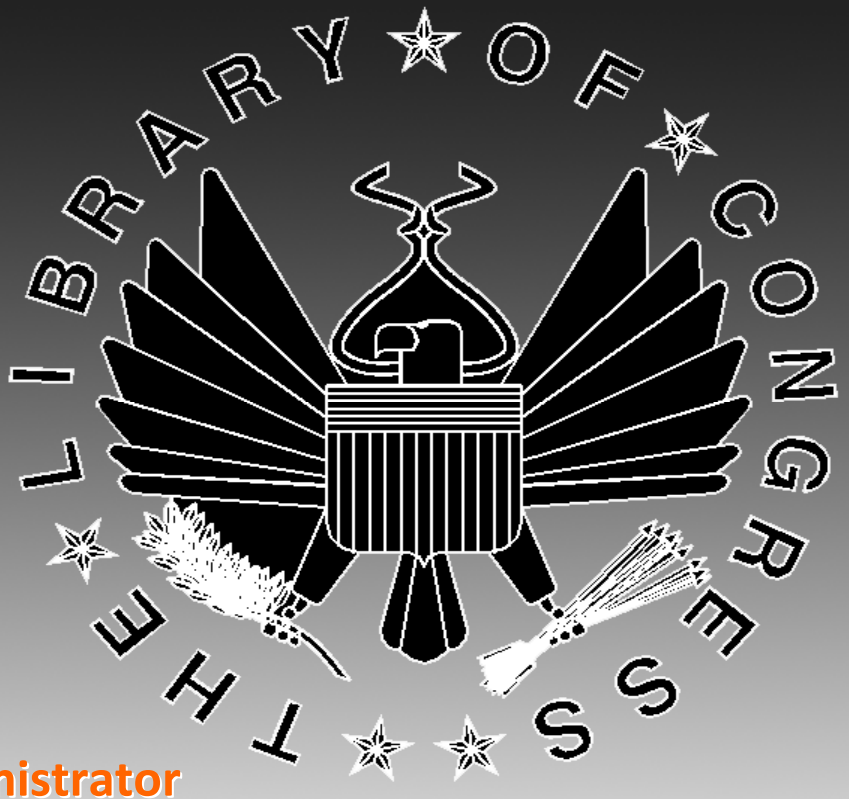
Future Challenges

- Film scanning:
  - Real time 4k film scanners with non-bayered imagers
  - Test 8k film scanning for 35mm
  - Develop mass-migration capabilities for our 255 million feet of film

Future Challenges

- Finding enough equipment to keep the migrations going
- Growing the Digital Repository into the exabyte realm…and beyond?
  - Not that far away!
- Developing the knowledge and training needed to make sure the 2-4 GENERATIONS of employees working on this project are adequately trained with proper documentation
- Getting the most bang for the bucks spent
- Funding (IE finding the bucks to spend)

Thank You!



James Snyder
Senior Systems Administrator
National Audio Visual Conservation Center
Culpeper, Virginia
jsny@loc.gov 202-707-7097