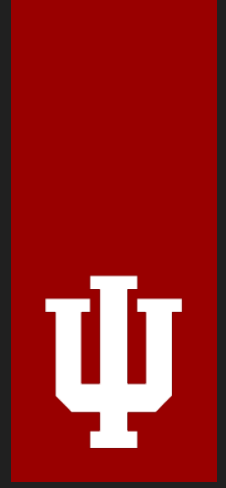


IU MDPI Status Update (Now with Film!)

Brian Wheeler / Senior System Engineer / Indiana University Libraries
bdwheele@indiana.edu

Media Digitization and Preservation Initiative

- "… to digitize, preserve, and make universally available by IU's bicentennial – subject to copyright or other legal restrictions – all of the time-based media objects on all campuses of IU judged important by experts" – IU President McRobbie, 10/2013
  - IU's Bicentennial is 2020
  - Film was delayed as part of a second phase.

Brief Timeline

- Phase I: Audio/Video (10/2013)
  - Audio began 06/2015. Video began 11/2015.
  - Under budget! Project extended from 280,000 to 325,000 objects
  - Bulk A/V digitization to end 2019Q1
- Phase II: Film (~9/2016)
  - Film digitization began 10/2017 and will end 2020Q4
  - Goal is 25,000 – 33,000 reels

A/V Digitization

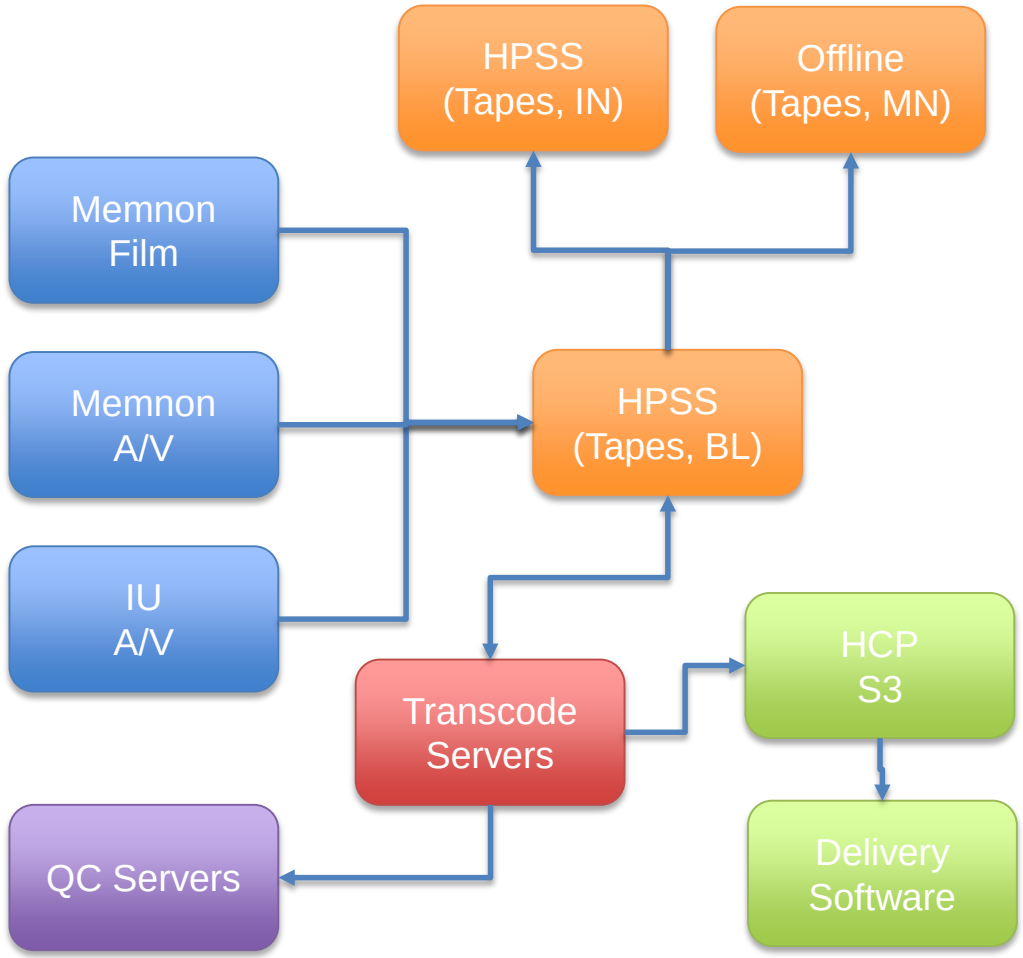
- A/V objects are digitized by Memnon Archiving Services and IU Staff
  - Memnon handles “bulk” digitization: 8-12TB per day
  - IU deals with unique and fragile items: ~1TB per day
- Roughly 10% are quality checked by humans

Film Digitization

- Memnon digitizes all of the film
  - 2x Laser Graphics Scanners
  - 2 shifts, 6 days per week
- Estimated: 16 content hours/day @ 26T
  - Reality: 8-12 content hours/day
  - Transfers 3-4 days per week @ 22 – 27T

Storage Architecture

- Digitizers (blue)
  - Send masters to HPSS (orange)
  - HPSS automatically makes 2nd copy in Indianapolis
  - Offline copy of tapes is made manually and sent to Minnesota
- Transcoders (red):
  - Retrieve masters
  - QC and create derivatives
  - Store derivatives in HPSS, a few go to the QC Servers
- QC Staff (purple):
  - Check portion of objects
  - Remainder get automatically passed
- Transcoders (red):
  - Collect metadata and send to delivery software (Avalon)
  - Derivatives to Hitachi Content Platform S3
- Delivery Software (green):
  - Content served to users



Challenges

Bandwidth

- Film is a lot bigger than A/V
  - 6T/hour (4K) or 1.5T/hour (2K) for film – “only” 64G/hour for video
- Designed for 35-40TB/day – we hit that regularly
  - 40Gbs backbone between HPSS, digitizers, and transcoders
  - Separate link for transcoders to HPSS vs rest of transcoder traffic

HPSS

- Tape writes are not verified
  - All content must be purged from the disk cache and re-read from tape
  - Tape reads and writes must be synchronized to avoid shoe-shining
- Staging woes – files are not grabbed in tape order
  - Created a staging process to feed HPSS stage requests in best order
- No programmable interface available (to us)
  - Had to wrap the `hsi` command and read output

Film Preservation Format

- BagIt tarball used for preservation
  - DPX frames, WAV audio, and metadata
  - Verifying content of tarball is time consuming
    - Wrote multithreaded verification tool to check in less than 25% runtime
- Preservation too large to move around for Human QC purposes
  - Memnon creates a ProRes proxy from the preservation DPXs

Data Protection

- HPSS is managed by the same team in both Bloomington and Indy
  - A rogue actor could wipe out years worth of work
  - Offsite copy of tapes created manually and sent to somewhere in Minnesota
  - MDPI files hardlinked to a no-access directory within HPSS so link count > 1
- Waiting for next version of HPSS to use media validation vs file fixity

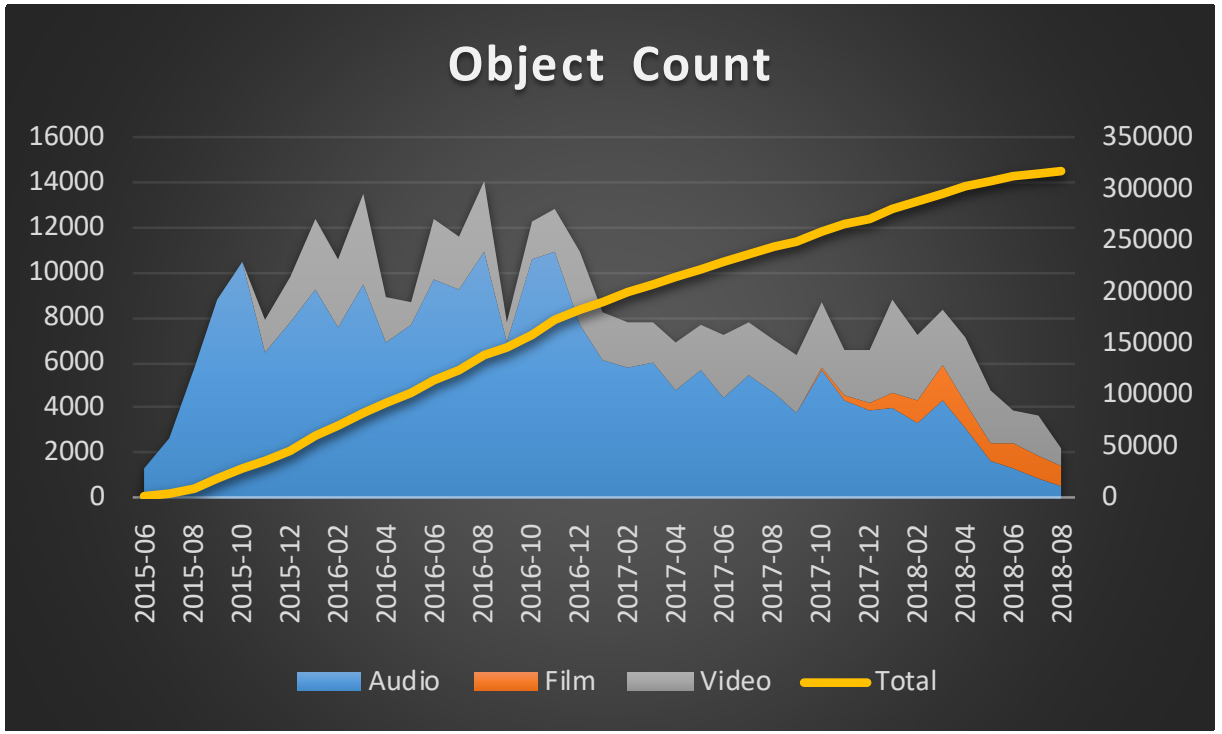
Vendor

- Scanner Hardware Problems + QC Bottlenecks
  - Failed objects and re-scans lead to unpredictable data flow rates
  - Vendor delaying delivery to verify content and then flooding the system
- Malformed Data Files
  - Despite extensive checking on everything else, the “BagIt-Version” line in bagit.txt was actually “BagIt-version” so we have to download, fix, and upload >2PB of data.

Where we are now…

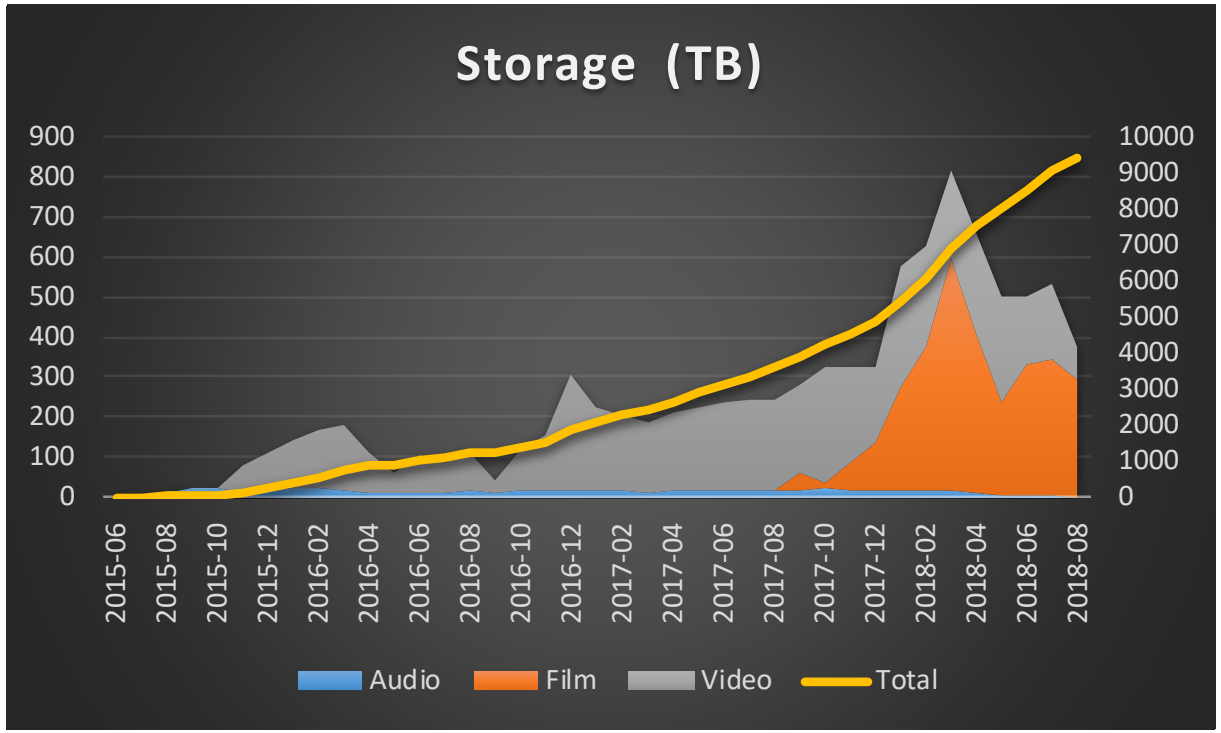
Object Counts

- More than 310,000 objects
  - Average > 8,000 per month
  - Almost 270,000 hours of content
- Totals curve is levelling off…
  - Audio is winding down
  - Video is focusing on long VHS



Object Storage

- Around 9PB of Storage
  - 1 year for 1st PB – now every 2 months!
  - Peak of > 800TB in 3/2018
  - Curve should remain about the same
- ~9 million files
  - Masters, derivatives, and metadata
- 27PB Estimated when finished



Future

- Digitization continues
  - Bulk A/V continues for next 6 months
  - IU will continue to digitize A/V for the foreseeable future
  - Film digitization will continue for another 27 months
- Long term preservation
  - We’re investigating long-term out-of-region copy options
- Data cleanup
  - Some multipart objects have missing or undigitized parts
- Provide easy access to media files:
  - Derivatives on demand for collection managers or researchers

Thank you!

- Media Digitization and Preservation Initiative
  https://mdpi.iu.edu
- Avalon Media System
  http://avalonmediasystem.org

INDIANA UNIVERSITY
FULFILLING *the* PROMISE