



**LIBRARIES**

# **MDPI Wrap Up**

Brian Wheeler / Lead System Architect / IU Libraries  
bdwhee@iu.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 was 2020
  - Digitization extended into 2021 (A/V) and 2022 (Film)
- Materials digitized by two sources:
  - External vendor located on campus for bulk digitization
  - In-house staff for delicate/unique materials



# Final Counts

- 372K physical objects digitized
  - 280K A/V estimated, extended to 325K – 340K actual
  - 25K - 30K films estimated, 32K actual
- 308K hours of content
  - Includes “dead air” – digitized to end of media
- Generated ~23PB of content (per copy)
  - Online copies in Bloomington and Indianapolis
  - Offline copy in Minnesota

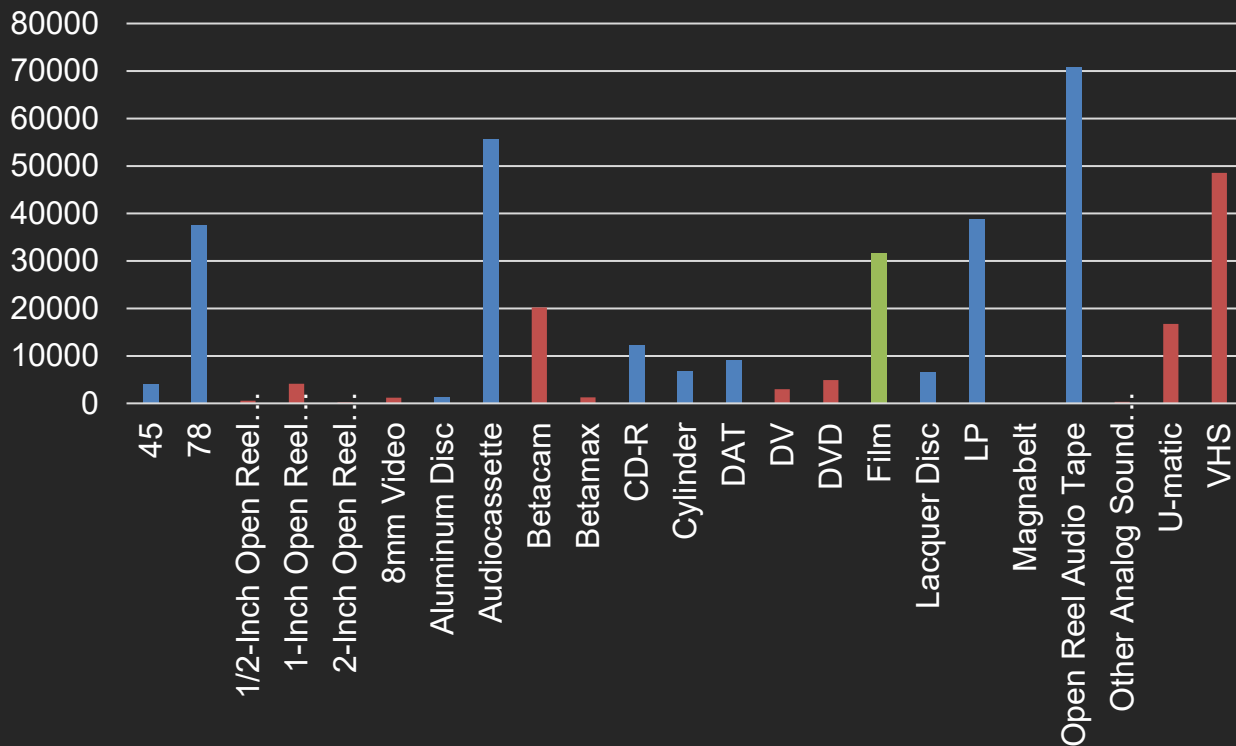


# Throughput

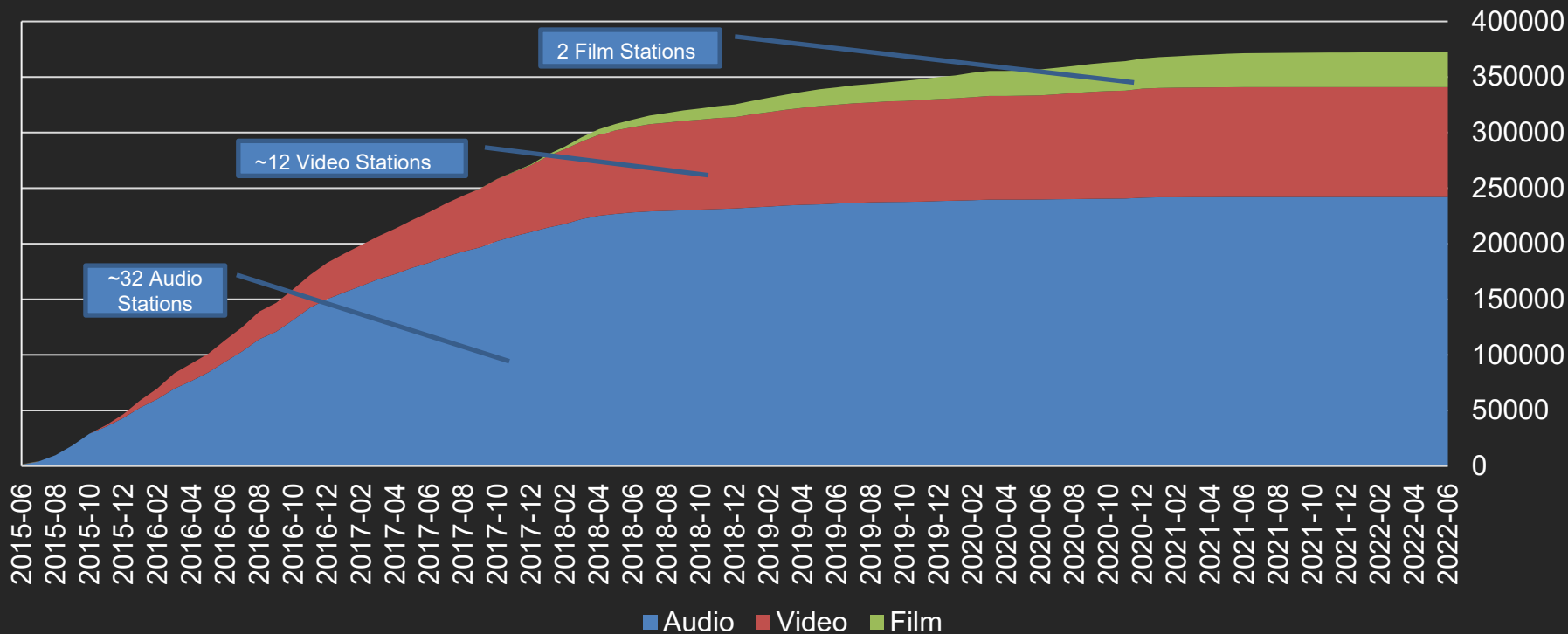
- Original Maximum Estimates:
  - 10TB/day for A/V
  - 35TB/day for A/V + Film
- Actual
  - 61 days of  $\geq 35$ TB/day
  - 30 days of  $\geq 40$ TB/day
  - 1 day of  $\geq 50$ TB/day
  - Overall average: 10.7TB/day
    - Includes holidays, weekends, project wind down



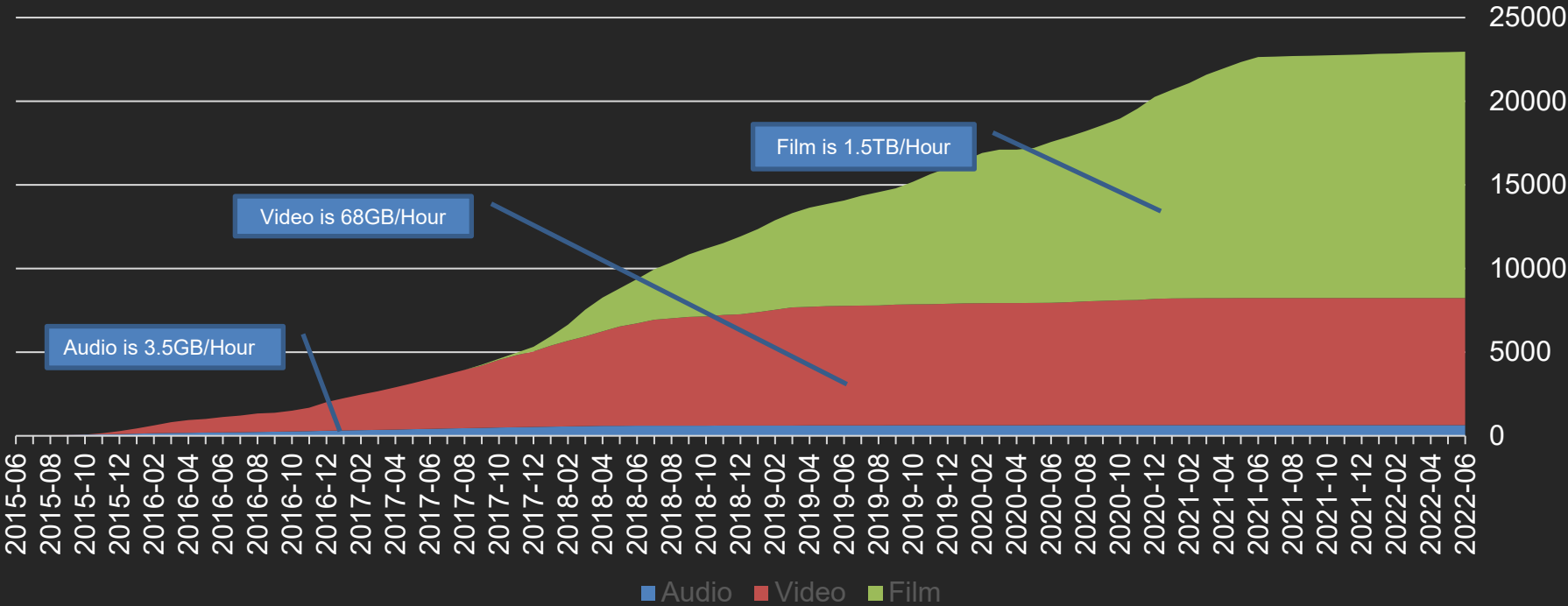
## Object Count per Format



# Objects Digitized by Type



# Usage (TB)



# Post-Digitization Cleanup



# Identifying Unused Content

- Large quantity of objects on tape which are not needed:
  - Accidentally resent – Vendor would sometimes resend objects
  - Digital file issues – Format/encoding issues, corrupt files, etc.
  - Content issues – Bad digitize, color correction, frame rate, etc.
- Redigitization of films with bad audio
  - Received 100s of films with poor audio quality from scan hardware
  - Reprocessing fix promised by scanner vendor fell through
- Total unused content:
  - 14K objects, split evenly between film and A/V
  - 3.6PB storage, 97% film



# Removing unused content

- Preparation
  - Scripts generated lists of unused objects and reason for removal
  - Verified by collection managers. Around 50 unused objects retained
- Offsite copy not modified
  - Safety net in case something went terribly wrong
  - But it would be expensive to recall tapes
- Permanent Removal from HPSS Tapes
  - Without a doubt the most terrifying process of my professional life
    - Three days of nausea
  - Everything validated correctly at the end
    - Final storage size 19.5PB per copy



# Reduce tape footprint

- Tapes have lots of holes or aren't full
  - Tapes in the middle of our workflow: many overwritten/erased files
  - Unused object cleanup created many more holes
  - Utilized multiple pools during MDPI – lots of half full tapes
- University IT Tape Library Migration (in progress)
  - Over-the-wire migration: smaller = faster
  - Tape data written without holes: fewer tapes = cheaper
  - New tape media (across both campuses)
    - JC (300), JD (5058) => JE (~2300)





# Future Directions

# Archival Management Software

- Investigating two commercial products for archival management
  - Libnova and Preservica
  - Early in the evaluation process
  - Hope to have RFP requirements by November
- Continue using our tape libraries for primary and secondary copies
  - Looking at an S3 frontend to archival data (vs HPSS)
  - Likely a virtual library to keep archival data separate
- Treat MDPI data the same as our other data
- Would like an in-place ingest for our tape data, but that's unlikely



# Out-of-region copy

- Current offsite third copy of MDPI tapes
  - Already paid for, minimal fees for storage
  - Includes the unused data that was deleted from onsite copies
  - We will need to retain a JD-compatible drive onsite
- Future third copy
  - Investigating cloud-based storage
  - “Copy of last resort”
  - Likely very little fixity checking due to costs



# Enhance metadata and content identification

- Too many MDPI objects without usable metadata:
  - “Untitled” – 784 objects, “No Title” – 1102 objects
  - Some with IDs that only refer to a spreadsheet somewhere
- Lots of dead air in many objects (digitized to end of media)
  - No idea how much “real” content we have
- We’re investigating tools to generate better metadata:
  - Speech to Text, Video OCR, Silence/black screen detection, etc.
  - While not perfect, they can be a start
  - Audiovisual Metadata Platform grant wrapping up



Closing thought:

MDPI has been the most challenging project I've ever been involved with.  
I had a blast.



**INDIANA UNIVERSITY**



# Days per Terabyte Processed

