First Author's Notes:

The first author's field notes are not included in this dataset because he played more of a supervisory role during the data collection phase of the project. Instead, he reviewed all of his co-authors' field notes, checking for consistency and accuracy.

Second Author's Notes:

The second author's field notes are not included in this dataset because she wrote her sections of the paper resulting from this research based on the tacit knowledge she gained while employed by MDPI for two years prior to joining our research team. She also reviewed all of her co-authors' field notes, checking for consistency and accuracy.

Third Author's Notes:

On October 6th IU IT administrative officer spoke in our digital curation class about her work at the MDPI. The IU IT administrative officer focused primarily on the MDPI at a higher, administrative level, providing a complementary presentation to the more technical talk given earlier in the semester. But the IU IT administrative officer did balance the administrative with personal stories from her work at the MDPI.

Mostly due to the previous presentation in our digital curation class, I was already fairly well informed on the purpose and function of the MDPI. The IU IT administrative officer reiterated the goal of the MDPI, its motivation, and general organization. The IU IT administrative officer gave some interesting additional information about the nascence of the project. IU researchers conducted a comprehensive survey of the at-risk audio visual materials held at IU and presented a plan for preserving them. This report was recognized by congress as a reference model for other large scale media digitization projects.

The IU IT administrative officer also gave some insight into the hardware that the MDPI employs. At this point, all objects digitized by the MDPI end up on tape storage in the Scholarly Data Archive (SDA). The SDA has extensive (upwards of 4 PT I believe) magnetic tape storage here at IUB and also in Indianapolis. The tapes are replaced approximately every 4 years. The tape system is old but effective. Apparently, IU's partner in the MDPI, Sony, has recently developed a technique to store huge amounts of data on linked Blu-ray disks and this might be used in the future. There has been talk of transition to storage with the DPN but currently the DPN is unable to ingest the amount of information that the MDPI outputs.

The IU IT administrative officer brought up one major issue with the MDPI project that also came up during the previous MDPI presentation, a distinct lack of metadata. The IU IT administrative officer illustrated this problem with a personal example with the MDPI. The IU IT administrative officer heard tell of a wonderful and rare performance by the illustrious Hoagy Carmichael, one filled with wonderful show tunes and witty banter. But when The IU IT administrative officer tried to search for this Hoagy Carmichael performance she was unable to find it. Eventually, the performance was located thanks to an astute technician who remembered the exact vinyl and side on which he or her had heard the recording. There was no metadata associated with the recording identifying it as Hoagy Carmichael and thus the recording was not discoverable.

This anecdote illustrates an issue with metadata creation in the MDPI project. At this point, all the metadata associated with the created digital objects is technical. This includes format, physical materials, condition, etc. Clearly, this has and will lead to issues in terms of discoverability. From what I could gather, the physical label on the object is included in the metadata. It stands to reason that some goodly proportion of the digital objects will have a

content identifying label. The MDPI is digitizing over 300,000 pieces so going through even 5% of these pieces to determine the content would be a monumental task. But the MDPI could be doing more to aid archivists further down the line. The IU IT administrative officer said that for most (if not all, it was not totally clear) of the objects digitized an operator listens to some portion for quality control. This would seem an opportune time to note something of the content. Even if the operators are unable to identify the content exactly, notes on instrumentation and type of venue it would save time later.

Provide a brief summary of what the guest speaker(s) discussed.

- What, if anything, did you learn that you did not know before as a result of their presentation?

MDPI has been recognized as a reference model by congress in terms of mass digitization

- Describe the way(s) in which their work relates to digital curation.

- What, if anything, surprised you about their work?

- What, if any, challenges to performing digital curation work did you notice or the guest speakers identify and how, in your opinion, might those challenges be addressed?

Metadata

- Your general reactions to what the guest speakers discussed.

- All reflective essays are due 24 hours before the class period following each guest speaker's presentation (i.e., Wednesdays at 9:30am)

IU administrative officer:

- MDPI is an essential partnership

- A reference model in terms of mass digitization
- Digitize 300,000 audio-visual objects
    - Decided worthy of preserving
    - Did all of these get surveyed to determine worth?
- Over halfway done now, should be halfway done by 2018
    - But we'll be adding film so the project won't be over
- Foundational work published in 2010
    - How do we get high level support for such a large project?
    - IU scholars did a survey to get the attention of the provost
    - The lib of congress recognized this work as a model for surveys of media
- Digital was chosen over analog
    - Easier to do, distribute, discovery, copy, replicate...lots of benefits
- We produce 7tb a day, way too much for the DPN to handle
- Curation project
    - Jacobs school
    - Orson wells
    - Lacquer discs
    - Wax cylinders
    - Women in newspapers
    - U-matic videotape
- There are 80 units in the library guiding choice/preparation
    - Every step of the way is tracked through unique barcodes on objects
- Factory line has automated checks that notify operators when unexpected changes occur
- Files end up in SDA on tape
    - But IU's partner Sony has a new way to use Blu-ray that they might switch to
- Almost no one interacts directly with the master file, there are copies of varying resolutions for everyone
- Laura heard about a really cool Hoagy Carmichael performance
    - But she couldn't find it because the metadata did not mention Hoagy anywhere
    - The vinyl was labeled with a random title, and so was all of the generated metadata
    - It was only found because some random tech remembered listening to it and what the disk was
- Is there a time span for the keeping the original materials?
    - We're sentimental
    - It's up to the unit
- How are you moving all this data?
    - Building of mdpi is physically connected by 20g lines directly between buildings
    - For motion picture we plan to double network capacity
- Do you have plans to clean the digitized files

    - Just to preserve them in their current state
- Lots of people have intellectual property at stake so we have to know to whom we can release these files
- When is metadata made?
    - That's more of a post process for the archivist
    - Even in the factory production people are listening to them
    - Conceivably metadata could be generated at these times

Fourth Author's Notes:

Week 7 Class Notes: Guest Speaker from MDPI

Preservation Action, etc. (MDPI overview – Ref. model-mass digitization)

- Digitize 300,000 a/v formats before media degrade
- 2009 Media Pres conducted survey 569k holding -- published media survey to much acclaim
- Digitize for long term preservation, discoverability/access, multiple formats, replication, (out of region storage, relationships with other institutions, large files HAPI Trust)
- Notable items from people like: Joshua Bell at 14, Orson Wells, Wax Cylinders 1888 recordings, U-matic video
- MDPI Process:
    1. Choice/Preparation – inventory, prioritize, Batch, Que (staging media)
    2. Digitization Factory – Memnon and IU Operation (preparation, cleaning, digitization, QA)
    3. Access and Discovery to Scholarship – metadata, Rights issues, technical aspects, IMPAC (Archiving Media)
- Produce 7 TB a day
- They have their own large network to transfer – capacity is 20 gigs – will increase to 40 gigs for film
- No enhancements- just extract recording
- Baking/cleaning materials
- Digitization
- Quality assurance
- Some digitization is automated
- Sometimes a person monitors entire process
- Digital storage – check integrity of files
- They have a large 'digital archive' (servers) – robotic system where all of MDPI and IU are stored
- Have checksums

- They transfer materials to new tapes every 5-7 years
    - Investigating use of Blu-ray instead
- Create derivatives of files: preservation master remains untouched, scholars use access copy
- Refresh storage tech
- File formats continue to evolve so need to make sure the format has a long term future and keep aware of trends to update or transfer if needed
- When digitizing they are purists – get best reproduction and can enhance later if needed
- Keep original media in archive if possible
- MDPI found previously unknown Hoagie Carmichael recording
- They will eventually clean up files – especially for film
- The huge dark archive hasn't been released yet (not available)
- Have a copyright librarian working on rights issues
- Have to enhance metadata (by librarians) and research access
- Do structured metadata
- Librarians and archivists will enhance metadata and provide content metadata
- Now have basic metadata - enhance later post digitization
- Operators listen to the recording during digitization
- How did the project/collaboration start??
    - Business model
    - Relationships with campus leaders
    - IT and libraries have a good relationship
    - Archivists and music school started – should have had libraries early
    - Cost was a factor that helped push to create partnership and use of industrial digitization
    - Test period with Memnon then it started

Week 10 Class Notes Oct 27, 2016 – MDPI lecture

- Media preservation at scale
    - Digitization 280,000 AV recordings in next 4 years
    - Degradation and obsolescence

    - Analog matter is degrading and eventually can't use materials or hear/see videos
    - Old formats old equipment, high cost for repair, need expertise for repair
    - Eventually it will be too expensive to repair and only have 10-15 years
    - If did 1:1 preservation workflow = 58 to 120 years
    - Avpreserve.com why media preservation can't wait
- Survey= more than 560,000 audio, visual, film
- Held at 80 different units on campus
- 44% 248,000 materials are unique or rare
- Meeting the Challenge of Digital Preservation
- MDPI Created in 2013 – digitally preserve all significant av material
- MEMNON
- Funded by UITS, libraries, and provost
- Manage
- Digitization
- 30 grad students = SMART TEAM
- Library
- IT
- Software
- Access
- Communications
- Staff holding media
- Digital Strategy
    - Memnon does most digitization – multiple at one time
    - Parallel transfer (industrial scale) workflow
    - IU – 1:1 work flow for fragile formats/problems
    - 6.5 PB data
    - Project file formats
        - Standard audio format BlUF 24/96
        - Video has no standard
    - Prep-pre workflow digitization
    - Feed Memnon

    - Create 9 TB per day
    - SMART Team gathers metadata, transfers materials, barcodes
    - Physical object database (POD)
    - Tracks materials being preserved
    - IU Audio
        - Digitize 7000 field cylinders, lacquer disks, mixed speed tapes, wire recordings
        - Sound directions 1:1 workflows
    - Video
        - Hi 8/8 mm Betamax
        - Problem VHS, umatic, Betacam SP
    - Theory of Constraints
        - Work with bottlenecks
        - SCRUM methodology
        - Backlog
    - Quality Control
        - 914 files per day
        - 861 audio files/53 video files
        - Peak video = 191 files per day
        - Random sample
            - Visual aural metadata inspection
            - Automated QC tools
        - Mass storage = IU Scholarly Data archive bit storage at IUB and at IUPUI
        - Fedora preservation system
        - Access copies
        - File and metadata managed in Hydra DAM2
    - HydraDam2 be able to maintain access
        - Preservation repository- NEH funded WGBH partnership
        - Fedora software = access repository and Avalon media
- Memnon (Sony company)
- 1. Archivists and Libraries = TLC

- 2. MDPI = digitization (part of IT team)
- 3. Storage = long term hydra dam (storage)
    - Access and discoverability in Avalon
    - Lead by IU
    - Trying to keep all tools used open access
    - But time works against this process
- 4. Degralescence
- 5. Factory – digitize reliable and well working machine
- 6. Time
- Best Practices
    - IU leading with media digitization
    - MDPI key person is a leader in field – wrote standards
    - Using FFE1 files while LOC used JPEG 2000 which are now stuck that way
    - Others followed FFE1
    - People, programming, procedure
    - Memnon production asset management system
        - Documentation
        - And communication
        - All XML that goes with material
    - Procedures constantly being updated
    - Quality assurance check- assess batch profile and technical specifications at each step
    - Quality assurance is cheaper – fix at each step in the process
    - Quality assurance at the end is expensive
    - When beginning they have manifest, bins, barcodes
    - Bake= Systematic ultrasonic leaning
    - Playback and Ingest= clean tapes and machine
    - Quality control at end = should be ok because of QA
        - fail/pass/pass but check (get someone else to look it over)
    - Good Factory – total quality management always seeking to improve
    - Memnon has been able to update and learn new procedures for new materials

    - FFE1 file format made IU/Memnon a leader
    - Productivity improves with collaboration

MDPI Tour

- Servers
- Packages =
    - Preservation file Lg
    - Mezzanine file (compressed, high quality) to share and copy
    - Access file (Avalon and stream on web)
    - QC Tools open source files
    - XML files describe package
    - MD5 files used to check the package was sent successfully
- Floors are sound proof walls for studios/repurposing
- Security for intellectual property so it can't be stolen or lost
- NOA ingest software
- Sound – wav files; access=MP4 – all output files are open source if possible
- Migration = time based media – open source file formats
- Access = metadata, online access, use (rights issues)
- Long term storage
    - 3 copies
    - Infrastructure
    - Preservation over time
    - Object management system

Fifth Author's Notes:

Page 1.

Lecture Notes Oct. 6, 2016

MDPI: guest speaker

IU reference model

Preservation: hard to focus only here, interrelated

Actions: cleaning, validating, metadata

High level support: survey and then suggestions

Digital vs. analog; digital = long-term preservation, discoverability and access; multiple formats; replication

Wax cylinders

Staging, by libraries and archives; digitization= preparing, cleaning, preserving

Pure extraction

Tapes: 0-10 years; Blu-rays-longer lives

Write 2 maser files; derivative files

Original goes back to unit; they decide

Enhancing and very special pieces; choses by archives

Metadata enhancement of dark archives (e.g. copyrights)

Basic metadata is required

Collaborative relationships between IT and libraries

Emulation as Solution to digital longevity?

W/smaller institutions, utilized partnerships and vendors

Page 2.
Lecture Notes Oct. 20, 2016

Guest Lecturer from IU Libraries: Bit curator in a virtual machine

Working with archives – born Digital Preservation Lap

With New Media: - intentions of creators? Fully represent object; the way the media is structured

Disk imaging (5 TB out of F, image would take all 5); no reason to preserve materiality

Write blockers – can't change anything;

** checking for viruses ASAP, not connecting to Internet when possible?

Folders: received Created package; reading for ingest; move to long term storage;

With some instances, can use other software, depends on problems

** ethics, privacy issues; Donor agreements, only accessible on 1 computer for now

MDPI NOTES

Media Digitization and Preservation Initiative (preservation at scale)

Mike Case, Director of Technical Operations

Objective: 280k audio and video recordings over 4 years (2 left)

WHY?

Degralescence: degradation and obsolescence

Wax cylinders, lacquer disks, all analog and physical digital media objects actively degrading

Obsolescence – formats, equipment, repair parts, playback repair expertise, rods suppliers

Less than 15 years; 2 headed monster = too expensive or impossible

IASA Journal no 44; Jan 2015: The Gathering Storm: Why Media Preservation Can't Wait

More than 560 k audio video, film objects;

About 80 units on campus: libraries, archives, departments, a theater etc.

44% of holdings are unique or rare;

Media Preservation Working Group – Task Force – Advising Board

2013 State of the University Announcement: Preserving all audio and video by 2020 – university wide;

Funded by President, Provost, up for research, UITS, libraries

Colleagues around the world including NARA

Memnon- parallel transfers, industrial scale workflows

IU: fragile formats and problem items

6.5 PB in 4 years

Audio preservation Master- BWF, 29/96; Video Preservation Master: FFV1/matroska

Page 3.

Memnon:

Brought experience to help IU

More like a factory than a studio

Over 2,700 files created during peak days

Quality assurance throughout: every step

FFVI/Matroska recordings

Programming, procedures, manifests, and people

production asset management (PAM) system

3 Steps of Production:

1. Holdings: libraries, archives.

2. Memnon and IU – digitization

3. Storage, access, and preservation

**Open sound; Memnon as factory=best practices

1 A – machine - $60k; even modern (ish) formats falling apart

Mismanaged tapes

IU Leading the Case; 15 mil. Project

Memnon started in Brussels, just opened to other institutions

Programing:

Production Management System (PAM)

Barcodes, few keyboards and clutter

Copy of recording, but decemination;

* common language; efficient communication

Procedures:

Check products in constantly updated lock in – new procedures and training

Adjusting tapes and machines (MD5)

Quality control at end

Pass, fail, or Pass but Check (PBC)

People

Brussels – involved

Students and faculty at IU

Page 4.

Smart Team:

Feeding the Beast: 9 TB per day – 320 hours digitized, 616 objects, 2700 files

Physical object database – POD

IU- Audio: 7000 wax cylinders (1893-1930); lacquer disks, mixed speed tapes, wire recordings, …

IU Video: Hi 8 / 8 mm, Betamax, ½ EIAJ, and problem items

Theory of Constraints principles

Quality control – random sample (visual, audio metadata inspection)

Peak – 914 files per day

Digitized media mass storage, Fedora Preservation System, Avalon Media System, Data Center (built to withstand an F5 tornado)

Post processing Workflow = create derivatives, QC

Bit Storage (both Bloomington and Indianapolis, duel written)

Hydra DAM2 Fedora Software in a Hydra Framework

Progress to date (15 years in) over 150k recordings

For Memnon see other notes

Sixth Author's Notes:

Reflections for September 21

Provide a brief summary of what the guest speakers discussed.

A metadata analyst, and a senior systems engineer, who are both at IU, spoke about a range of data projects they have worked together on during their time at IU. They spoke briefly about their backgrounds. The Metadata analyst/librarian came from formal training in art history and the systems engineer from computer science. Most of their talk time was taken with presenting and demonstrating the need for human checks on computer conversion data systems.

What, if anything, did you learn that you did not know before as a result of their presentation?

I did not know the IU project was so much in isolation from other universities. Somehow, I thought we were more in connection with other music schools collecting and combining collections in our data storage facility. Maybe it just didn't come through in the presentation.

Describe the ways in which their work relates to digital curation.

Their presentation focused on the "create and receive" components but we spent our time in a demonstration of the ingest and preservation action portion of our lifecycle model. From the descriptions of their work it sounded more like the specific work that they did had more to do with oversight of systems. The Metadata analyst seemed to be concerned with descriptors and cataloguing of the various kinds of data that the MDPI and other projects housed in the university collections held. The systems engineer seemed most focused on solutions to problems of architecture for systems management for these collections. That said, both of them would be concerned with all aspects of the lifecycle model, if not overtly, at least tangentially.

What, if anything, surprised you about their work?

How much was learned on the job, and that more of their work now is focused on migration of existing collections rather than acquisition of collections.

What, if any, challenges to performing digital curation work did you notice or the guest speakers identify, and how in your opinion might those challenges be addressed?

The Systems Engineer said he had to make the systems work "all the time" and "had to make it up as we go along" which must be a terrible pressure. He offered that he drew on his own experience with Lego robotics, and warned that disks are fragile… "Make sure you backup, ALWAYS.

Your general reactions to what the guest speakers discussed.

So far it seems to me every speaker had to have a high tolerance for stress and low expectation of time away from work. It was reassuring to hear them say, however, that they believed the appraisal process should not be a single person's decision, and that before moving forward about the best one can do is know all the options and then select the "best" one known at that time.

Seventh Author's Notes:

IU IT Administrator

MDPI - Media digitization and preservation initiative IU

IU Associate VP for Information Technology and Deputy Chief Information Order Executive Director

MDPI:

- Over 500,000 objects -> 300,000 AV need to be preserved in 5 years
- Envy of peers in Big10
- Digitization was the route to preserve AV resources (originally believed in transferring)
    - Long term preservation is easily
    - Geographically dispersed
    - Larger storage
    - Discoverability and Access efficiently
    - Produce in multiple formats (e.g. access for low speed connection users)

Challenge:

- How do you get high level support?
- IU: College of Arts and Sciences, Library Provost
- Present solutions to get support
- Out of region storage: AP Trust, digital preservation network, model doesn't scale for IU
- What is the file format and what path do you have to convert/migrate in the future?

Edward S. Curtis Collection

Model:

- Choice/Preparation: Inventory, prioritizing, batching, queuing (library)
- Digitization factor: Memnon/IU operation
- Access and discovery to scholarship: metadata, rights issues, technical aspects, IMPAC

Memnon purchased by Sony, Memnon est. a factory in Bloomington, originally from Brussels

No enhancements or corrections, only extract best possible file

Scholarly data archive: record tape shelves, high capacity, IU's first massive data storage, checksum process

2 preservation masters, users use a different file with lower resolution to get to user quickly and no extra software needed for access

Access file is used by users

Moving/transfer files: private line to move information, motion picture will double network to 40 gigs

Open source file format, uncompressed

Purest: clean resource and play it through reproduction that is what the file should be (e.g. if videos have color problems, leave them)

Dark archive that hasn't been released, needs to have rights and permissions check

Needs to have corresponding metadata attached

Didn't have archivists/librarians at the table at the start of the project but are now at the table and making improvements with their improvement

When persuading people to spend money, you have to make numbers realistic to something people can relate too (e.g. staff positions, scholarships, etc.)

Avalon is built on community open source model
LOC promotes open source for access vs. proprietary
Investors (e.g. members in community model) can shape the program to how you want, and all other people can use it but can't shape it to their needs