AMPlifying AV
Next Steps for the Audiovisual Metadata Platform
CNI Spring 2022
March 29, 2022

Jon Dunn / Indiana University / @jwdunn
Shawn Averkamp / AVP / @saverkamp
Project website: https://go.iu.edu/amppd
Outline

● Background
● Phase 2 accomplishments and lessons learned
● Phase 3 goals and progress
● Next steps
Background
The Challenge

- Growing AV collections
  - Legacy formats
  - Explosion of born-digital
- Increased expectations for access
- Insufficient metadata
- Limited resources for cataloging
The Opportunity

● Mass digitization approach extended to AV
  ○ “Digitize first”
  ○ e.g. IU’s Media Digitization and Preservation Initiative (MDPI): mdpi.iu.edu

● Emergence and continued improvement of machine learning and other automated tools

● How can we leverage the best of automated tools and human expertise in flexible and configurable ways?
  ○ Diverse collections demand diverse workflows
AMP: The Vision

- Open source software platform to support metadata creation for AV collections
- Design and execute workflows combining automated and human steps
- Integrate multiple “Metadata Generation Mechanisms” (MGMs)
  - Automated, manual
  - Local, HPC, cloud
- Delivery of metadata to variety of target systems, e.g. online access systems (Avalon, Aviary), library catalogs, etc.
Work to Date

- **2017-2018: AMP Planning Project (Phase 1)**
  - Developed initial AMP concept
  - Convened technical architecture planning workshop
  - White paper: *AMP Planning Project: Progress Report and Next Steps*

- **2018-2021: AMPPD - AMP Pilot Development (Phase 2)**
  - Built proof-of-concept AMP system
  - Pilot tested using two collections from IU and one from NYPL
  - White paper: *AMPPD Final Project Report*

- **2021-2022: AMP Phase 3 (current)**
  - Additional system development
  - Packaging and deployment work
  - Testing with additional collections from IU and NYPL
AMPPD (Phase 2)
Phase 2 Accomplishments

● Developed AMP application
  ○ Web-based user interface: Java, Vue.js, React.js
  ○ Workflow management and execution: Galaxy

● Evaluated and implemented MGMs (Metadata Generation Mechanisms)
  ○ Automated and Human MGMs
  ○ Evaluation criteria

● Tested with 100 hours of audio and video from each of three collections
  ○ Indiana University Archives
  ○ Indiana University Cook Music Library
  ○ NYPL Gay Men’s Health Crisis collection
<table>
<thead>
<tr>
<th>Date</th>
<th>Submitter</th>
<th>Collection</th>
<th>External Source</th>
<th>External ID</th>
<th>Item</th>
<th>Primaryfile</th>
<th>Workflow</th>
<th>Step</th>
<th>Output</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>2022/03/24 01:58:40 pm</td>
<td><a href="mailto:amp@iu.edu">amp@iu.edu</a></td>
<td>Admin Collection</td>
<td></td>
<td>Lunchroom Manners</td>
<td>Lunchroom Manners</td>
<td>Lunchroom Manners</td>
<td>MW - Transcription and NER with HMGM for NER</td>
<td>transcript_to_webvtt</td>
<td>web_vtt</td>
<td>Complete</td>
</tr>
<tr>
<td>2022/03/24 01:55:03 pm</td>
<td><a href="mailto:amp@iu.edu">amp@iu.edu</a></td>
<td>Admin Collection</td>
<td></td>
<td>Lunchroom Manners</td>
<td>Lunchroom Manners</td>
<td>Lunchroom Manners</td>
<td>MW - Transcription and NER with HMGM for NER</td>
<td>ner_to_csv</td>
<td>hmgm_ner</td>
<td>Scheduled</td>
</tr>
</tbody>
</table>
Chancellor Wells, thank you for inviting us out this afternoon to visit with you. Uh, the community of course is very closely tied with Indiana University. And to me, I guess you are Mr. Indiana University. We like this afternoon, if we could to maybe visit with you concerning the history uh your comments about Bloomington. I know you came here as a student in 1921. Tell us what bloomington was like in 1921.

Booming in 1921 was of course much smaller than it is now, it was much more typical of an Indiana uh county seat town than it is now. Uh the growth since never fails to amaze me that I drive out around the growth of population, the growth of oh, of the extent of the city, the the building and so on. Uh but it was it was a thriving, is uh um active county seat center in those days with the close relationship between the university on the one hand and the downtown leadership particularly on the other, I found it very well it was a beautiful place then as it is now uh there were certain features of that period that especially attractive. The the the fifth Avenue or Kirkwood coming from town to town was arts over with big trees all the way out, it was rather rural and even greener than it is now possible. The old campus here was not quite a stick with trees as it is now 50 or 60 years made a difference. They the trees grow like all the rest of us and they all they not only grow taller, they grow around there, don’t we? All.

The students in those days did they, were they involved as much in the community as we see the students in the last 15 years where they active in political causes and the social causes.

And I don’t think they were in as much involved with social causes as they became after World War Two. When the Colonel Shoemaker was dean of students, he got started these various projects in which students would undertake to sponsor uh uh philanthropic and humanitarian activities downtown. Uh it started with, he got, it started with the Amp houses, I recall exchanging their bell week to Help week and the work.

fraternities and

sororities. And then they

were active on the
Phase 2 MGMs Implemented

- **Speech-to-text:**
  - Kaldi (OSS, HPC)
  - AWS Transcribe (CCL)

- **Named entity recognition:**
  - SpaCy (OSS)
  - AWS Comprehend (CCL)

- **Video OCR:**
  - Tesseract (OSS)
  - MS Azure Video Indexer (CCL)

- **Segmentation:**
  - PySceneDetect (OSS)
  - MS Azure Video Indexer (CCL)
  - INA Speech Segmenter (OSS, HPC)
  - Applause Detection (AMP)

- **Human MGMs:**
  - BBC Transcript Editor (OSS)
  - Audio Timeliner for NER editing (OSS)

- **Other:**
  - Dlib face_recognition.py (OSS)
  - FFmpeg (OSS)
  - Vocabulary tagging (AMP)
  - Gentle forced alignment (OSS)

OSS=Open source software  
CCL=Commercial cloud  
HPC=High Performance Computing  
AMP=AMP-developed
MGM Evaluation Criteria

- Accuracy
- Input formats
- Output formats
- Growth rate
- Processing time
- Computing resources
- Social impact / ethical considerations
- Cost
- Support
- Integration capabilities
- Training
Phase 2 Lessons Learned

● Challenges with proprietary tools
  ○ Unpredictability, undocumented behavior
  ○ “Black box” aspects of process
  ○ Lack of clarity in terms of use
  ○ Opt-in vs. opt-out for use of data for product improvement
  ○ …but for certain tasks they work really well!

● Tools for language more robust and available than those for music
  ○ Speech-to-text, NER vs. Music genre, instrumentation detection
  ○ Many music use cases require training
  ○ Opportunity for training data sets from libraries, collaboration with music IR community
Phase 2 Lessons Learned (continued)

- Use of existing tools/models vs. training of new models
  - Applause detection for music concert segmentation
  - Facial recognition - ethical considerations

- Technical implementation
  - Unpredictable wait times for batch-oriented HPC jobs
  - Integrating Human MGMs into Galaxy workflows

- Librarian/archivist engagement
  - Lots of excitement about potential
  - More difficult to think about practical implementation
AMP Phase 3
AMP Phase 3

Funded by a new grant from the Mellon Foundation
July 2021 – December 2022

Focus on:

- System Robustness and Resilience
- Packaging, Deployment and Documentation
- UX Evaluation and Improvement
- MGM Evaluation Module
- Collection Testing
UX Evaluation and Improvement

Collection Partners

Indiana University
● Archives of African American Music and Culture
● Archives of Traditional Music
● Black Film Center & Archive
● IU Libraries Moving Image Archive
● University Archives

New York Public Library
● Research Division (Rights, Archives, Metadata)
● Schomburg Center for Research in Black Culture (Moving Image and Recorded Sound Division)
Phase 3 Feature Development

- Workflow creation
- Batch upload
- Improved navigation of collections
- “Intermediary” files as MGM input
Phase 3 Feature Development

- HMGM tool improvements
  - BBC Transcript Editor debugging
  - Addition of video player to Timeliner (editing NER, Audio segmentation, others)
- Independence of HMGMs from AMP workflows

NER Editor

Lunchroom Manners
Named Entity Recognition for Lunchroom Manners

Source: https://amo.dlib.indiana.edu/review/mon/14/files/47/media
Engaging with Collections Managers

- System demos
- Hands-on training
- All-team meetings
- Targeted discussions
- Focus groups
Use cases for AMP

- **Archivists** want to efficiently accession collections
- **Catalogers**, **archivists** and **metadata librarians** want to describe media or enrich metadata
- **Rights clearance staff** want to provide appropriate levels of access
- **Researchers** want minimally processed media to be further described on-demand
System Requirements for File Upload

- **Archivists** want to efficiently accession collections
- **Catalogers, archivists** and **metadata librarians** want to describe media or enrich metadata
- **Rights clearance staff** want to provide appropriate levels of access
- **Researchers** want minimally processed media to be further described on-demand

Flexibilities in data model and API to accommodate:
- Single item, batch, and API upload
- 0 or more external identifiers
Data Output Formats

A “contact sheet” that can be generated from shot detection, video OCR, face recognition, and other MGM outputs.
User Feedback – Surfacing New Issues

PySceneDetect (shot detection)
contact sheet

Too many shot transitions detected in areas with no content
User Feedback – Surfacing New Issues

00:28:36 --> 00:28:39
<v spk_0> T. I. U or call monday through friday

00:28:39 --> 00:28:40
<v spk_4> during normal business hours.

00:29:05 --> 00:29:05
<v spk_4> Okay.

00:29:14 --> 00:29:14
<v spk_4> Okay.

00:29:30 --> 00:29:34
<v spk_4> Okay. Okay.

00:29:51 --> 00:29:54
<v spk_4> Okay. Okay. Okay.

00:30:01 --> 00:31:06
<v spk_4> Okay. Okay thank you. Okay. Okay.

00:31:27 --> 00:31:27
<v spk_4> Okay.

00:31:39 --> 00:31:49

00:32:01 --> 00:32:08
<v spk_4> Okay. Okay.

00:32:37 --> 00:32:44
<v spk_4> Okay. Okay.

00:32:58 --> 00:33:04
<v spk_4> Okay. Okay.

Amazon Transcribe (speech-to-text transcription) WebVTT file
Words transcribed from segments of silence
MGM Evaluation Module

**Speech-to-Text**
Speech-to-text transcription (also known as automatic speech recognition, or ASR) is the recognition of spoken language in an audio stream and conversion to text.

**Video OCR**
With supporting text below as a natural lead-in to additional content.

**Audio segmentation**
With supporting text below as a natural lead-in to additional content.

**Applause Detection**
With supporting text below as a natural lead-in to additional content.

**Named Entity Recognition**
With supporting text below as a natural lead-in to additional content.

**Face Recognition**
With supporting text below as a natural lead-in to additional content.

(work in progress wireframes)
Evaluation – Ground Truth Testing

MGM Evaluation

Speech-to-Text
Speech-to-text transcription (also known as automatic speech recognition, or ASR) is the recognition of spoken language in an audio stream and conversion to text.

→ Confluence documentation Sample Use Cases

Select a New Test

STT Word Error Rate Test
STT Accuracy Test

STT Word Error Rate Test
Word error rate measures speech-to-text accuracy by comparing the generated transcript against a ground truth transcription and calculating the number of errors (substitutions, additions, and deletions) as the cost of restoring the output word sequence to the original input sequence. Scores are measured as percentages, with lower scores representing high accuracy (low word error rate). Scores may exceed 100%, especially if the STT engine produced many insertions. Related to WER, character error rate (CER) is similar to WER, but based on characters, the word information loss (WIL) measures the proportion of word information lost in a transcription, and word information processed (WIP) measures the inverse of WIL.

Download the Ground Truth Template

(work in progress wireframes)
Evaluation – Visualization Interface

(work in progress wireframes)
## Qualitative Evaluation

<table>
<thead>
<tr>
<th>Ground Truth</th>
<th>MGM Output</th>
<th>Error Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>its</td>
<td>its</td>
<td></td>
</tr>
<tr>
<td>my</td>
<td>like</td>
<td>substitution</td>
</tr>
<tr>
<td>very</td>
<td>very</td>
<td></td>
</tr>
<tr>
<td>distinct</td>
<td>distinct</td>
<td></td>
</tr>
<tr>
<td>pleasure</td>
<td>pleasure</td>
<td></td>
</tr>
<tr>
<td>to</td>
<td>to</td>
<td></td>
</tr>
<tr>
<td>welcome</td>
<td>welcome</td>
<td></td>
</tr>
<tr>
<td>you</td>
<td>you</td>
<td></td>
</tr>
<tr>
<td>all</td>
<td>all</td>
<td></td>
</tr>
<tr>
<td>to</td>
<td>over</td>
<td>substitution</td>
</tr>
<tr>
<td>the</td>
<td>the</td>
<td></td>
</tr>
<tr>
<td>second</td>
<td>second</td>
<td></td>
</tr>
<tr>
<td>patten</td>
<td>**</td>
<td>deletion</td>
</tr>
<tr>
<td>lecture</td>
<td>lecture</td>
<td></td>
</tr>
<tr>
<td>of</td>
<td>off</td>
<td>substitution</td>
</tr>
<tr>
<td>this</td>
<td>this</td>
<td></td>
</tr>
</tbody>
</table>

*Ground truth comparison view of Amazon Transcribe speech-to-text output*
Packaging

Multi-tier deployment approach

- Tier 1: Direct install to OS with installation scripts and detailed documentation
- Tier 2: Install component-level Docker containers
- Tier 3: Scripts and configurations for orchestration of container deployment

https://github.com/AudiovisualMetadataPlatform (work in progress)
AMP Future

Stay Tuned!
AMP Future

- Release and adoption of AMP for local and cloud installation
- Integrate additional MGMs and target systems
- Opportunities for training new models
- Support output as IIIF / Web Annotations
- More work around music use cases
- Continue to build community for AI/ML in libraries and cultural heritage (e.g. AI4LAM)
Thank you!

go.iu.edu/amppd

Jon Dunn: jwd@iu.edu

Shawn Averkamp: shawn@weareavp.com