12-10-2021

Audiovisual Metadata Platform Pilot Development (AMPPD), Final Project Report

Jon W. Dunn
Indiana University - Bloomington

Ying Feng
Indiana University - Bloomington

Juliet L. Hardesty
Indiana University - Bloomington

Brian Wheeler
Indiana University - Bloomington

Maria Whitaker
Indiana University - Bloomington

Audiovisual Metadata Platform Pilot Development (AMPPD)

Final Project Report

December 10, 2021

https://go.iu.edu/amppd

INDIANA UNIVERSITY

avp

New York Public Library

TEXAS
The University of Texas at Austin

Report Publication Date

December 10, 2021

Report Authors

Jon W. Dunn, Ying Feng, Juliet L. Hardesty, Brian Wheeler, Maria Whitaker, and Thomas Whittaker, Indiana University Libraries

Shawn Averkamp, Bertram Lyons, and Amy Rudersdorf, AVP

Tanya Clement and Liz Fischer, University of Texas at Austin Department of English

The authors wish to thank Rachael Kosinski and Patrick Sovereign for formatting and editing assistance.

Funding Acknowledgement

The work described in this report was made possible by a grant from the Andrew W. Mellon Foundation.

Table of Contents

PROBLEM STATEMENT

Libraries and archives hold massive collections of audiovisual recordings from a diverse range of timeframes, cultures, and contexts that are of great interest across many disciplines and communities.[1]

In recent years, increased concern over the longevity of physical audiovisual formats due to issues of media degradation and obsolescence,[2] combined with the decreasing cost of digital storage, have led institutions to embark on projects to digitize recordings for purposes of long-term preservation and improved access. Simultaneously, the growth of born-digital audiovisual content, which struggles with its own issues of stability and imminent obsolescence, has skyrocketed and continues to grow exponentially.

In 2010, the Council on Libraries and Information Resources (CLIR) and the Library of Congress reported in "The State of Recorded Sound Preservation in the United States: A National Legacy at Risk in the Digital Age" that the complexity of preserving and accessing physical audiovisual collections goes far beyond digital reformatting. This complexity, which includes factors such as the cost to digitize the originals and manage the digital surrogates, is evidenced by the fact that large audiovisual collections are not well represented in our national and international digital platforms. The relative paucity of audiovisual content in Europeana and the Digital Public Library of America is a testament to the difficulties that the GLAM (Galleries, Libraries, Archives, and Museums) community faces in creating access to their audiovisual collections. There has always been a desire for more audiovisual content in DPLA, even as staff members recognize the challenges and complexities this kind of content poses (massive storage requirements, lack of description, etc.). And, even though Europeana has made the collection of audiovisual content a focus of their work in recent years, as of February 2021, Europeana comprises 59% images and 38% text objects, but only 1% sound objects and 2% video objects.[3] DPLA is composed of 25% images and 54% text, with only 0.3% sound objects, and 0.6% video objects.[4]

Another reason, beyond cost, that audiovisual recordings are not widely accessible is the lack of sufficiently granular metadata to support identification, discovery, and use, or to support informed rights determination and access control and permissions decisions on the part of collections staff and users. Unlike textual materials—for which some degree of discovery may be provided through full-text indexing—without metadata detailing the content of the dynamic files, audiovisual materials cannot be located, used, and ultimately, understood.

---

[1] See for example, "Quantifying the Need: A Survey of Existing Sound Recordings in Collections in the United States," AVP and the Northeast Document Conservation Center, May 27, 2015, accessed June 25, 2021, https://www.weareavp.com/quantifying-the-need-a-survey-of-existing-sound-recordings-in-collections-in-the-united-states/

[2] Mike Casey, "Why Media Preservation Can't Wait: The Gathering Storm," *IASA Journal* 44 (2015): 14-22, accessed June 25, 2021, https://www.weareavp.com/mike-casey-why-media-preservation-cant-wait-the-gathering-storm/

[3] "Search," Europeana, accessed February 2021, http://www.europeana.eu/portal/en/search?q=.

[4] "Search," Digital Public Library of America, accessed February 2021, https://dp.la/search

Traditional approaches to metadata generation for audiovisual recordings rely almost entirely on manual description performed by experts—either by writing identifying information on a piece of physical media such as a tape cassette, typing bibliographic information into a database or spreadsheet, or creating collection- or series-level finding aids. The resource requirements and the lack of scalability to transfer even this limited information to a useful digital format that supports discovery presents an intractable problem. Lack of robust description stands in the way of access, ultimately resulting in the inability to derive full value from digitized and born-digital collections of audiovisual content, which in turn can lead to lack of interest, use, and potential loss of a collection entirely to obsolescence and media degradation.

PROJECT GOALS

Since October 2018, the IU Libraries—in collaboration with the University of Texas at Austin, information innovation company AVP,[5] and NYPL—have worked to help address these challenges through the creation of an open-source software platform known as AMP (Audiovisual Metadata Platform),[6] which is designed to enable more efficient generation of metadata to support discovery and use of digitized and born-digital audio and moving image collections. This project and the planning project that preceded it in 2017 have been generously supported by the Andrew W. Mellon Foundation, with substantial in-kind staff and computing contributions from IU.

The overarching goal of the Audiovisual Metadata Platform Pilot Development (AMPPD) project, which took place from October 2018 through June 2021, was to develop enough of the AMP system to be able to pilot test it using two audiovisual (AV) collections from IU and a third collection from NYPL. The project team has developed a software system that harnesses the Galaxy workflow engine,[7] originally developed for data processing workflows in computational genomics, to design and execute custom workflows for metadata and feature extraction from AV files.

As part of this work, the team also evaluated metadata generation mechanisms (MGMs) in eight different categories and selected, where possible, at least one open source and one commercial cloud solution within each category, including speech-to-text, named entity recognition, audio segmentation, video OCR, scene/shot detection, structured OCR of supplementary materials, known-person facial recognition, and applause detection. The software development team then created “wrappers” for each selected MGM to allow them to be plugged into workflows within Galaxy for execution through AMP. In addition to these automated MGMs, so-called “human MGMs” (HMGMs) were implemented to allow human intervention in workflows when necessary to perform actions such as correcting speech-to-text output, selecting desired terms from named entity recognition, and validating and adjusting the results of automated segmentation.

---

[5] AVP website, accessed June 28, 2021, https://www.weareavp.com/.
[6] Homepage, Audiovisual Metadata Platform Pilot Development (AMPPD) wiki, accessed June 25, 2021, https://go.iu.edu/amppd
[7] Homepage, Galaxy Community Hub, accessed June 23, 2021. https://galaxyproject.org/

Based on work and results so far, the project team has concluded that the approach taken in AMP is effective and scalable for generation of metadata for certain types of AV collections, particularly those that involve significant amounts of spoken word content. This includes lectures, events, and documentaries, along with oral history interviews and other ethnographic content.

Among the specific goals of the AMPPD project are those listed below, alongside the means by which they were accomplished:

Technical Architecture Goals

***Functionality*:** The architecture contains the system components necessary for collections staff to create workflows of MGMs, schedule those workflows, assign those workflows to specified sets of files, store the metadata that is generated, and publish the metadata that is generated.

***Configurability*:** The modular approach we are taking will allow for components to be updated over time as technologies advance. The ability to interface with different storage environments, import data from different source systems, and publish to different target systems all speak to the configurability of the architecture. This includes a fully implemented API.

***Ease of use*:** The User Interface Application (UIA) component speaks most directly to the ease of use. The UIA is intended to provide a non-expert a simple way of configuring and executing workflows from a palette of MGMs without being burdened by the complex architecture behind the UIA.

Flexibility in adapting to new workflows and MGM implementations: The ability to "plug in" MGMs and support an ecosystem of MGMs representing local, cloud, open source, closed source, free, paid, automated, and manual options provides a great deal of flexibility from the start and over time. This is achieved through the use of the Galaxy workflow engine; basically, all that Galaxy requires from any new tool is an XML file with the specification details of how to execute it.[8] This feature and Galaxy's type-checking at the time of workflow creation ensure that tools can be lined up correctly with respect to input requirements. Through this mechanism, we have integrated a variety of commercial and open-source automated and human MGMs, including AWS Transcribe and Kaldi for speech recognition, AWS Comprehend and spaCy for named entity recognition, BBC Transcript Editor for transcript correction, etc. As MGM technologies evolve, the AMP architecture will be able to incorporate these changes, allowing users to leverage new technology capabilities in their workflows.

MGM Goals

Identify and employ MGMs appropriate for use in AMPPD based on the following criteria:

- Accuracy

[8] "Galaxy Tool XML File," Galaxy Community Hub, accessed June 23, 2021. https://docs.galaxyproject.org/en/master/dev/schema.html

- Input formats
- Output formats
- Growth rate
- Processing time
- Computing resources
- Ethical considerations
- Cost
- Support
- Integration capabilities
- Training

Full descriptions of each are available later in this document.

BACKGROUND

PROJECT HISTORY

The AMPPD project was preceded by a 2017 planning workshop hosted by IU and resulting white paper as part of a Mellon-funded planning project to inform the design and development of AMP. The follow-up project, AMP Pilot Development (AMPPD), kicked off in late 2018 and wrapped up in June 2021. Funding for a third phase has recently been awarded by Mellon, which began in July 2021 and will run for 18 months.

AMP Planning Project Workshop

The 2017 AMP planning workshop was specifically focused on (1) determining the technical details necessary to build the platform and (2) bridging the gap between prior work of the project partners and future implementation. The workshop brought together individuals from within and outside the partner organizations, all of whom have relevant expertise and experience to assist the partners in analyzing the needs for the system and identifying the best technologies and approaches to building a functioning prototype. The workshop participants were:

- Adeel Ahmad, AVP (Former AMPPD Project Team Member)
- Kristian Allen, UCLA Library
- Jon Cameron, Indiana University
- Tanya Clement, University of Texas at Austin (AMPPD Project Team Member)
- Jon Dunn, Indiana University (AMPPD Project Team Member)
- Maria Esteva, Texas Advanced Computing Center, University of Texas at Austin
- Michael Giarlo, Stanford University
- Juliet Hardesty, Indiana University (AMPPD Project Team Member)
- Chris Lacinak, AVP (AMPPD Project Team Member)
- Brian McFee, Music and Audio Research Laboratory, New York University
- Scott Rife, Library of Congress
- Sadie Roosa, WGBH Media Library and Archives

- Amy Rudersdorf, AVP (AMPPD Project Team Member)
- Felix Saurbier, German National Library of Science and Technology
- Brian Wheeler, Indiana University (AMPPD Project Team Member)
- Maria Whitaker, Indiana University (AMPPD Project Team Member)

In the years leading up to this workshop, the project partners had embarked upon various initiatives investigating audiovisual description. In 2015, IU and AVP investigated models and developed a strategy for high-throughput description of audiovisual materials that are being digitized as part of IU's Media Digitization Preservation Initiative (MDPI).[9] AVP gathered information through interviews with collections staff at IU and users of MDPI content to understand whether metadata exists (it often does not), and if so, in which formats (video, audio, handwritten documents), applications (spreadsheets, databases), and/or structures (XML, CSV, TXT) it resides. Collections staff also identified optimal output formats and potential uses for the metadata and considered related rights and permissions issues for the digitized objects and their metadata. These interviews resulted in (a) the establishment of a set of metadata fields for optimized discovery of AV assets in IU's Media Collections Online[10] AV access system based on the open-source Avalon Media System[11] jointly developed by IU and Northwestern University, (b) identification of the metadata fields' value for discovery beyond Avalon, and (c) the values of those fields in the generation of other or subsequent metadata (e.g., general keywords can be analyzed to produce specific names, subject terms, and dates).

AVP then identified, through market research and interviews with developers of systems including Nexidia, Fraunhofer's AV Toolbox, Perfect Memory, and Apex, nearly thirty existing metadata generation mechanisms (MGMs) for populating the proposed metadata fields. These include, for example, AI/machine learning applications for natural language processing, facial recognition, legacy closed caption recovery, as well as human-generated metadata and OCR of images and transcription, which have the potential for capturing and producing metadata at a massive scale when unified in the modular AMP architecture.

AVP's initial research led to a proposal for an iterative approach to metadata capture, generation, and enhanced re-generation, wherein the full suite of envisioned MGMs would be deployed in three phases. In this model, first-phase MGMs would produce sets of data that could be analyzed by second and third-phase MGMs. By phase three, MGMs would begin to integrate various outputs from early processes to augment granular and topical description, ultimately increasing discoverability and usability. Throughout the three phases, AMP would act as the workflow engine, pushing data from one MGM to the next, as well as:

- serving as a decision engine, continuously evaluating results at all processing stages (e.g., MGMs, workflow processing) and routing data through workflows accordingly. For instance, identifying content as speech versus music and routing to the appropriate processing path,

[9] "Media Digitization & Preservation Initiative," Indiana University, accessed June 23, 2021, https://mdpi.iu.edu.
[10] Media Collections Online, Indiana University, accessed June 28, 2021, https://media.dlib.indiana.edu/.
[11] Homepage, Avalon Media System, accessed June 23, 2021, https://avalonmediasystem.org. *This project has been funded in part by grants from the Andrew W. Mellon Foundation and Institute of Museum and Library Services.*

- storing metadata for processing,
- providing a metadata warehouse for longer-term storage of all metadata generated, and,
- serving as a metadata source for target systems, such as Avalon (for the pilot phase) and Aviary,[12] that offer metadata management and/or discovery related to AV content.

As part of their initial study, AVP analyzed costs, staffing allocations, technology, and services required to implement AMP at IU. This project offered IU:

- an architecture and strategy for AMP,
- a realistic high-level view of the resources, staffing, etc., required to implement AMP, and
- the opportunity for vast improvements to discoverability of and access to their audiovisual collections.

The MDPI metadata strategy project, then, provided a strong foundation for the 2017 AMP workshop and planning project discussions, which resulted in a white paper[13] released in March 2018 that summarized the output of the workshop and planning project and recommended the next phase of work that led to the current AMPPD project.

AMP Pilot Development Project

This white paper presents the findings of the Audiovisual Metadata Platform Pilot Development (AMPPD) project, which has worked to enable more efficient generation of metadata to support discovery and use of digitized and born-digital audio and moving image collections. The project was originally planned to take place over a period of 27 months beginning on October 1, 2018, and through a no-cost extension, continued through June 30, 2021. Funding from the Mellon Foundation has been augmented through substantial in-kind staff contributions from Indiana University. The AMP system, built as part of the AMPPD project, enables the creation and execution of workflows that link together both automated and human analysis activities, and it has been tested against representative media sample sets from three specific collections, drawn from the collections of IU and NYPL, that contain different content types (e.g., music and spoken word, documentary and performance, from different time periods and with differing image and audio quality), media types, and metadata extraction requirements.

For many collections, when using the metadata that existed prior to AMPPD, discovery opportunities were extremely limited. A user from IU who might have searched for longtime IU President Herman B Wells using “Wells” or “HB Wells” to find a video in the library catalog would have then needed to watch the entire video to see (a) whether Wells appeared on it and (b) where in the video he appears. Today, with AMP and the MGMs that are utilized in the platform (audio and video transcription, scene detection, and facial recognition), users not only know if Wells is in a video, but exactly where in the video he appears, and what he says or what is said about him. When ethically applied (see Ethical Considerations section below), this could be a game changer for large AV collections that otherwise have very little description.

---

[12] Aviary website, AVP, accessed June 28, 2021, https://www.aviaryplatform.com/.

[13] Jon W. Dunn, Juliet L. Hardesty, Tanya Clement, Chris Lacinak, and Amy Rudersdorf, “Audiovisual Metadata Platform (AMP) Planning Project: Progress Report and Next Steps,” March 27, 2018. http://hdl.handle.net/2022/21982

Leveraging the metadata from AMP, for example, users are already able to conduct searches (with varying levels of results) such as:

- Take me to every point in a video interview with Herman B Wells where Herman B Wells mentions Eleanor Roosevelt on the subjects of Presidents' spouses and 20th-century leaders.
- Show me every video interview with Herman B Wells in the 1970s where the interviewer is Thomas D. Clark, and it was produced at WTIU Bloomington.
- Take me to every point in a video interview with Herman B Wells where Herman B Wells is on camera and talking about Midwest universities where there is no music present.

Uncovering the underlying opportunities for metadata capture and exposure of vast AV collections was a major motivation for this project. What is being developed is an intuitive system that is easy for non-developers and non-technical caretakers of collections to use. We are hopeful this will change the prospect for future access to hundreds of millions of hours of AV content and open up collections in meaningful ways, such as data and content analysis at scale, with description not only about the media, but also extracted from the content of the media files, leading to discovery capabilities currently only available for text-based content. By the end of the project, the project team aimed to maximize findability and usability of AV assets by making AMP available to IU and NYPL libraries and archives as an open-source software platform with documented APIs that allow flexible integration with each institutions' digital content ingest workflows and access systems, along with basic documentation for the system's use.[14]

PROJECT ORGANIZATION

AMPPD TEAM

The AMPPD project team was composed of staff from four organizations: Indiana University, New York Public Library, University of Texas at Austin, and information innovation company AVP. The project staff members were broken into teams that included "Development," "MGMs," "Collections," and "Administration." Some staff belonged to more than one team. The list of staff and AMPPD teams and roles are listed below:

Averkamp, Shawn (AVP) Lead, MGM team
Boolchandani, Vinita (IU) Developer, Development team
Cameron, Jon (IU) Collections team support
Clement, Tanya (UT) MGM team special advisor
Dunn, Jon (IU) Principal Investigator
Duryee, Alexander (NYPL) Collections team
Feng, Ying (IU) Lead developer, Development team
Fischer, Dan (AVP) Developer, Development team

[14] Development of more complete technical and user documentation for AMP is a component of the proposed Phase III work

Fischer, Liz (UT) MGM team
Hahn, Michelle (IU) Collections team
Hardesty, Juliet (IU) Lead, Collections team; MGM team
Kellams, Dina (IU) Collections team
Lyons, Bertram (AVP) Advisor
Marri, Naresh (IU) Developer, Development team
McAfoose, Sarah (IU) HMGM transcription team
Mellon, Mary (IU) Collections team
Peters, Chuck (IU) Collections team
Rubinow, Sara (NYPL) Collections team
Rudersdorf, Amy (AVP) Project manager, Admin team; Scrum master, Development team; MGM team
Salibi-Cripe, Laila (IU) HMGM transcription team
Shea, Caroline (AVP) Developer, Development team
Sovereign, Patrick (IU) Project support, Development team and general
Sutton, Jack (IU) Project support, general
Timko, Karen (IU) HMGM transcription team
Wheeler, Brian (IU) System architect, Development team
Whitaker, Maria (IU) Project Owner, Development team; Administrative team; MGM team
Whittaker, Thomas (IU) Advisor and HMGM transcription team lead; MGM team
Yolles, Melanie (NYPL) Collections team

ADVISORY BOARD

The advisory board met mid-way through the project to discuss and provide insights on decision points about which the AMPPD project staff sought input.

Bruns, Gerrit. Competence Center for non-textual Materials, German National Library of Science and Technology
Giarlo, Michael. Stanford University Library
Hunter, Caitlin. Recorded Sound Section, Library of Congress
Kaufman Davis, Casey. WGBH Media Library and Archives
McFee, Brian. McFee Music and Audio Research Laboratory, New York University
Pustejovsky, James. Brandeis University, Department of Computer Science
Van Dall, Dirk. BAMTECH

In addition, the AMPPD team gained support and input from the following two groups:

Fraunhofer Institute for Digital Media Technology (IDMT). This team provided development support and access to their relevant MGMs.

Aichroth, Patrick
Kühhirt, Uwe
Lukashevich, Hanna

Sieland, Marcel
Taenzer, Michael
Weigel, Christian

The AMPPD core team also met monthly to share updates and approaches with the similarly AI/machine learning-focused Mellon-funded project called **"Computational Linguistic tools for Multimedia Services (CLAMS)."** The CLAMS team members included individuals from the project's institutional lead—WGBH—and their development partner, the Brandeis University's Lab for Linguistics and Computation.

Cariani, Karen (WGBH)
Kaufman, Casey Davis (WGBH)
Lepczyk, Timothy (WGBH)
Lynch, Kelley (Brandeis University)
Pustejovsky, James (Brandeis University)
Rim, Kyeongmin (Brandeis University)
Verhagen, Marc (Brandeis University)

AMPPD TIMELINE

The AMPPD project was divided into milestones and within that, the Development team worked in sprints of two weeks. The entire team met monthly, while a smaller "core" team of stakeholders met every other week to ensure the project stayed on track. The Development team also met daily for sprint standups of 15 minutes each. The MGM team met weekly for the bulk of the project; as the project wound down in the spring of 2021, those meetings were moved to twice monthly. The Collections team met on an ad hoc basis as needs arose.

The Project Owner developed a project roadmap that laid out the milestones and goals of each. That document can be found in Appendix A.

In the first phase (of three) of the grant, the Development team focused on selection of tools (including Galaxy), development of the core platform and UI, and building wrappers for the MGMs that were analyzed, tested, and ultimately selected by the MGM team. Phase two saw further development of the platform and more MGMs, including support for workflows including human intervention steps (Human MGMs). Additionally, the Development team tested the utility of using high-performance computing for some MGM tools. The final phase saw the continuation of platform and UI development and more MGM wrappers for the Development team; wrapping up analysis and selection of MGMs and a move to focus on documenting and reporting out on the selection criteria for each. The Collections team became very active in the last two phases, testing the UI and data outputs. The team was surveyed about their experiences with both and that data appears later in this report.

By the end of the grant, AMPPD was a fully functioning platform, integrating 24 MGMs that could be used to analyze, describe, and document the audiovisual materials in IU and NYPL's collections.

To prioritize evaluation and selection of MGMs for implementation in the AMP platform, the MGM team first reviewed the outputs of the Planning Project Workshop to identify metadata fields of interest and categories of tools that could potentially supply that metadata for audiovisual materials. After compiling a list of these tools, the Collections team identified and prioritized use cases for metadata generation for each collection, then ranked MGM categories by frequency of occurrence across collections. As the project progressed, additional use cases were developed and relevant categories selected for evaluation. The MGM team evaluated the following categories of tools, in order of priority (italics indicate tools that were evaluated but not selected for implementation):

- Speech-to-text transcription
- Audio segmentation (speech, silence, music, noise)
- Entity recognition (natural language processing)
- Video OCR
- *Music genre classification*
- Shot detection
- *Structured OCR of print materials (music programs)*
- Applause detection
- *Music ensemble classification (audio)*
- Facial recognition
- Forced alignment (speech and text)

For each category of tools, the team searched for both proprietary and open-source tools to evaluate, using GitHub Awesome Lists,[15] popular GitHub repositories, and frequently mentioned tools from blog posts and online articles as resources. For proprietary tools, the team looked at Google, AWS, and Microsoft Azure for commercial offerings within each category. The project team felt it was important to select a mix of proprietary and nonproprietary tools to demonstrate the flexibility of AMP in accommodating both self-hosted applications and external web services and to show how tools from different sources could be combined in pipelines. Additionally, though proprietary tools can present ethical concerns through their black-box nature, the low cost, advertised higher accuracy, and lower barrier to implementation can make them a more accessible option for under-resourced institutions. The MGM team aimed to select one proprietary and one nonproprietary tool from each category for implementation, when possible.

Next, the MGM team compiled a list of criteria for evaluating tools within each category to provide a framework for the team to have productive discussions with the Collection and Development teams about which tools would fit their needs and why. Criteria were drawn from a

[15] "Awesome List," Github, accessed June 24, 2021, https://github.com/topics/awesome-list.

number of sources, notably the Principles for Accountable Algorithms and a Social Impact Statement for Algorithms,[16] which guided criteria around social impact and accuracy. These criteria are intended to be a reusable tool for anyone wishing to implement machine learning at their organization—different criteria may have different weight for a given organization or project. While accuracy, a common measure for evaluating a machine learning tool, is important, it is only one factor in determining if a tool will be a good fit for an organization or its use cases. Cost, social impact, and processing time were also highly important considerations for the AMP project.

Evaluation Criteria	Description
Accuracy	How does the MGM output compare to the expected value (or human-generated value)?
Input formats	File types, encodings, compressions, etc., allowed by the MGM. Assess the level of difficulty involved in converting your files to the formats required for the tool. How will this impact automation? Is anything lost in the conversion that could affect the accuracy of output?
Output formats	File types or data formats output by the MGM. Assess the level of difficulty involved in converting available output formats to the desired format. How will this impact automation?
Growth rate	Rate of increase of time and computing resources as volume/file size increases. Compare processing time between small, average, and large-sized files to estimate time required as scale increases. Is this feasible given the estimated contents of your project? Compare memory use between small, average, and large-sized files to estimate memory required as scale increases. Is this feasible given the estimated contents of your project?
Processing time	Time required for the MGM to process the file. How will processing time affect your production workflows? Can processing time be improved by optimizing computing hardware, software, or networks?
Computing resources	Amount of computing resources, including processing power, memory, network connections, and bandwidth required to process the file. How will computing resources affect your production workflows? Will you need to operate the MGM on other machines?
Social impact	The potential unintended consequences of an unmediated MGM's output. What are the possible unintended negative impacts that could come from the output of this MGM? What measures can be taken to mitigate them? See FAT/ML's Principles for Accountable Algorithms for more information: http://www.fatml.org/resources/principles-for-accountable-algorithms

[16] "Principles for Accountable Algorithms and a Social Impact Statement for Algorithms," Fairness, Accountability, and Transparency in Machine Learning, accessed June 24, 2021, https://www.fatml.org/resources/principles-for-accountable-algorithms.

Cost	The cost of the MGM, which could include paid services, file transfer, and computing costs if running in the cloud, or local hardware and staff costs.
Support	Available human support, documentation, or logs output by the MGM which can help with learning or troubleshooting the MGM.
Integration capabilities	The ability of an MGM to fit into a workflow design or technical infrastructure or the ability to supply functionality for other computational needs, such as a speech-to-text tool that also provides segmentation and speaker diarization.
Training	Whether or not a model should be trained to utilize the MGM. Consider the costs, time, and social impact of training a model or using a model out of the box.

After a cursory evaluation of tools within each category, the team selected a shortlist of tools to test accuracy on a sample of collection content. For most MGMs, these samples were:

- (NYPL) Women & AIDS Teach-in: Single-camera video footage of a speaker event.
- (NYPL) Day of Desperation: Video footage (with video quality issues) of a protest with visible signs and many overlapping voices.
- (IU Archives) Little 500: A compilation of video footage from field day and indoor events at IU set to music with title slides between events.
- (IU Music Library) West Side Story: Single-camera video recording of a stage musical.
- (IU Archives) Student-Admin Forum: Single-camera video recording of a student forum in an auditorium.
- (IU Archives) Astin Patten Lecture: Single-camera video recording of an auditorium lecture. Includes text projected by overhead projector. Black and white.

The items chosen from each collection were representative of video for that collection and provided different qualities of video and audio for testing the shortlist of tools. NYPL's items showed amateur video footage of multiple scenes or events along with words on screen in the form of protest signs. IU Archives' items showed a variety of video content: one of a lecture series with a single speaker and words on screen in the form of transparency slides; a student forum in black and white video with an audio format where people on stage are speaking into microphones clearly and then audience members are speaking or shouting from their seats; and a compiled video with sound overlay that showed a variety of events spliced together. The IU Music Library video item was typical of video items in their digitized performance collection, with a single camera shot showing a staged performance from a distance. These items offered many different kinds of content to detect but also were typical representatives of digitized analog formats that could present challenges for the tools being tested.

For music-focused MGMs, a mix of commercial CDs and recordings of IU recitals from the IU Music Library supplemented the above list. For testing facial recognition, videos known to contain test subjects (former IU President Herman B Wells, and former IU Vice President Charlie Nelms) were used.

The MGM team installed open-source tools on local machines for testing and accessed most proprietary web services through IU organizational accounts. Testing with some tools halted at the installation stage when documentation was too lacking to successfully run the tool on samples.

For most categories of tools, the MGM team created ground-truth data for each of the samples tested and calculated accuracy scores (usually precision, recall, and F1[17]) for each. Ground-truth generation for speech-to-text transcription was outsourced to 3Play Media, but the rest of the ground-truth data was created manually using Google Sheets, Google Docs, or similar online office tools. Ground truth was typically exported as CSV and compared against MGM results that were similarly converted from JSON to CSV format. One challenge for the MGM team was the incredible amount of time and effort needed to write the conversion scripts from each MGM tool's original data output format to a common data structure for ground-truth comparison. The next phase of work will address this challenge by building these conversions for many tools into the platform to enable any users of AMP to supply ground-truth data for their own collections and easily test accuracy before choosing to use the tool.

In addition to making a quantitative assessment of MGM accuracy, the team also looked for ways to visualize results for qualitative assessment. This usually involved loading MGM outputs and ground-truth data into Google Sheets in a way that would allow the team to use filters and sorting to visually review the results. The team engaged the collections staff in reviewing the data, to get a sense of both the usefulness of the MGM outputs and the usefulness of the visualization format. Often the usefulness of the outputs could not easily be determined without presenting the data in a format or tool that allowed collections staff to navigate and review a high volume of data within the context of the media itself. Additionally, many tools produced a large amount of "noise" such that, even if useful data was included in the output, filtering it from false positives or reducing the level of granularity of data was difficult to do with the tools available. For example, video OCR tools captured text in every frame of a video, which was both redundant, because of text that stayed on the screen for many frames, and noisy, because of errors caused by illegible or moving text. While the output may have had high recall (i.e. the MGM found most of the correct words), it also generated too much noise for a human to feasibly ignore.

For some MGM categories, certain output formats proved to be somewhat useful to collections staff for review, but there remains more work to be done in this area of human review mechanisms to bridge the gap between MGM output and integration into library metadata. One

[17] Ground truth data is the ideal output of a machine learning algorithm. This data is compared to the actual machine learning output to calculate measures of accuracy. Precision indicates how closely a machine learning prediction aligns with the ground truth (i.e. there is a low number of false positives). Recall indicates how well the machine learning prediction identifies all ground truth instances. (i.e. there is a high number of matches, or "true positives"). F1 is the harmonic mean between precision and recall, considered a balance between the two.

output format the team developed shows some promise for certain use cases. The “contact sheet” is a series of frame images extracted from a video to represent data points in an MGM output. This format was designed during the testing of shot detection (detection of a series of frames captured by the same camera representing a continuous action or focus within a scene) in support of a use case for copyright determination, where a user needs to identify any visual presence of intellectual property, such as artwork on a wall, to assign appropriate permissions or restrictions on use of the video. By extracting the center frame of every shot, the user could view all of the frames simultaneously on the contact sheet to review all visual content in the video. Upon reviewing the contact sheets, the collections staff determined this format would lend itself to other use cases, such as getting a high-level overview of the content or absence of content on an unprocessed video or verifying classifications made by a facial recognition MGM. Collections staff also suggested generating contact sheets based on certain intervals as alternative to shot detection, which sometimes generated too much noise to be useful. Because of this collaboration between the MGM team and Collections team, these variations on contact sheet generation are now available in the AMP platform.

Though the primary goal of the MGM team was to make informed recommendations on which tools to include in the AMP pilot platform, the byproducts of this process may be valuable to future users of AMP or anyone wishing to test and select machine learning tools at their organization. For each category of MGM evaluated, the team produced cursory evaluations of tools according to the criteria mentioned above, implementation scripts for running the tools on the samples, scripts for comparing MGM outputs against ground truth data, and a quantitative and qualitative analysis of the tools reviewed. These artifacts are shared publicly through the project wiki (Confluence),[18] with code shared on the project’s organizational GitHub page.[19]

HUMAN MGMs

As stated in the project proposal, for the greatest success, automated mechanisms must work in concert with human labor managed by a recursive and reflexive workflow engine that supports an ecosystem of open-source and proprietary tools and services—local and cloud-based. Thus, the concept of human metadata generation mechanisms (or “HMGMs”) became part of the workflow schedule. The AMPPD team initially identified three areas where HMGMs could be implemented to enhance or refine outputs from automated workflows: speech-to-text transcript correction, named entity revision, and audio segmentation refinement. By the end of the grant period, HMGMs for transcript correction and named entity revision were fully incorporated in AMP. The refinement of audio segmentation data was not explored during the pilot, as it was determined that this work would be undertaken outside of AMP (e.g., in a target system such as Avalon).

[18] :”Documentation,” AMP wiki, accessed June 28, 2021, https://wiki.dlib.indiana.edu/display/AMP/Documentation.
[19]“AMP: Audiovisual Metadata Platform,” Github, accessed June 22, 2021, https://github.com/AudiovisualMetadataPlatform.

Since the Collections team identified speech-to-text transcription as a priority for investigation, particular attention was given to the transcript correction HMGM. In addition to being useful as a surrogate for the content, a corrected transcript with accurate timestamps facilitates navigation and feeds into other downstream automated MGMs (e.g., natural language processing). While producing human-corrected transcripts can be time intensive due to the need to listen, verify, and update text, it was hoped that by starting with the output from automated speech-to-text we could present a feasible alternative to a fully automated or fully manual process for transcript generation.

Existing cataloging staff from Indiana University were engaged for transcript correction rather than hiring temporary staff. Professional catalogers are familiar and comfortable with work that requires a high degree of accuracy and a close attention to detail. External factors also contributed to this decision. In particular, all work would need to be done remotely, with no in-person interaction (the bulk of this work took place during the COVID-19 pandemic). The perceived challenge of recruiting, hiring, training, and managing temporary staff in a fully remote environment was a significant factor in the decision to use existing staff.

Recognizing that there was considerable interest in testing fully automated MGM workflows, as well as workflows with human intervention, only a representative sample of content was sent through the HMGM workflow. As such, the files that went through the transcript correction/named entity revision workflow were selected based on length of content rather than nature of content. This resulted in a semi-random selection of a combined fifteen hours of content from IU Archives and NYPL (IU Music Library content was not selected) with a mix of video and audio-only materials.

The following is a start-to-finish account of the transcript correction/named entity revision HMGM workflow. A more complete description of the workflow tools and other tools used to perform the work and their integration with AMP can be found in the Technology section of this paper.

Initiating the Transcript-NER-HMGM workflow in the workflow engine (Galaxy) led to the creation of a ticket in the task management tool (Jira[20]) for transcript correction. Each ticket included a link to the speech-to-text-produced transcript in the integrated transcript editor (BBC Transcript Editor). The HMGM manager would then assign these tickets to the HMGM staff. The staff would open the file in the transcript editor and correct the transcript as needed, recording the amount of time spent on each file on the respective Jira ticket. Once finished, HMGM staff would inform the HMGM manager that the file was ready for review. The HMGM manager would perform a final review of the corrected transcript in the transcript editor and click the “complete” button to submit the final transcript. Upon completion, the workflow would automatically close the Jira ticket, designating the task complete, and would initiate the next step in the workflow, sending the corrected transcript through the natural language processing MGM. A new Jira

[20] https://www.atlassian.com/software/jira, accessed November 2, 2021.

ticket would then be created for named entity revision. In much the same manner as transcript correction, the HMGM manager and staff would then complete named entity revision in the integrated tool (Avalon Timeliner[21]).

TECHNOLOGY

System Architecture

The system architecture for AMP was largely informed by the output of the platform architecture workshop conducted during the project's planning phase in 2017-2018. The first few months of work by the technical team during the AMPPD phase were dedicated to researching open questions and refining the architecture, choosing tools where necessary.

The AMP architecture involves a combination of existing software components and new components developed by the AMP development team. A diagram of AMP's current technical architecture is shown in Figure 1:

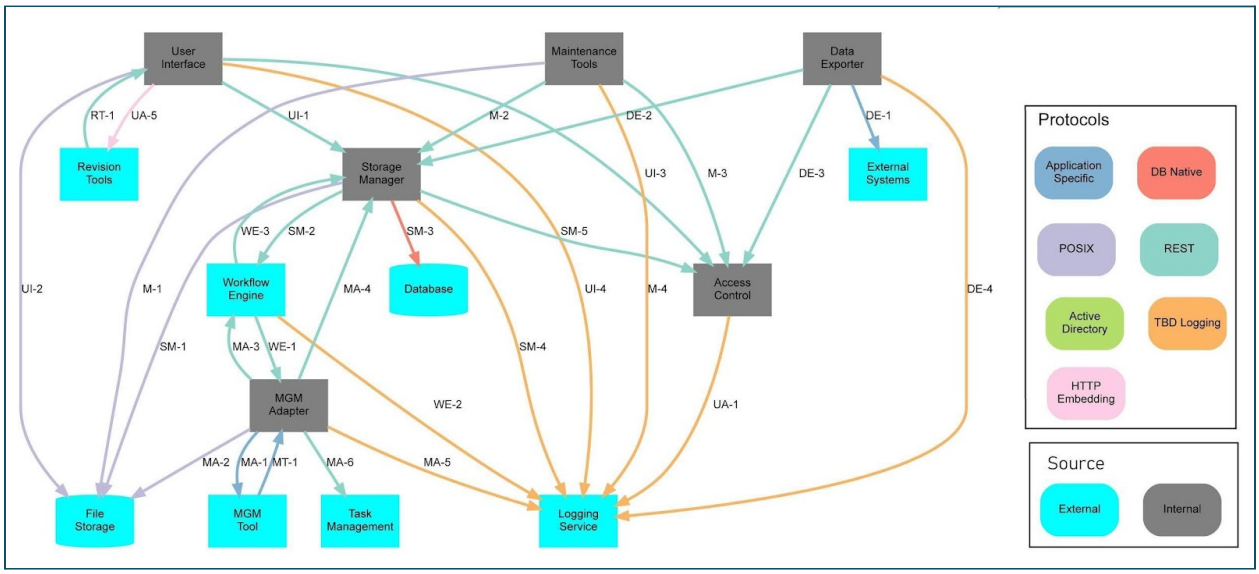


Figure 1. AMP architecture

AMP application

The modules identified as internal in Figure 1 are responsible for the orchestration of the various other components of the AMP application. Other than the Access Control module, which has not

[21] "Timeliner," AMP Github site, accessed June 28, 2021, https://github.com/AudiovisualMetadataPlatform/timeliner.

been worked on, the other internal modules are written in Java, Python, Vue.js, or React.js, and are available on the project's GitHub repository.[22]

The AMP user starts by ingesting into the AMP system the AV content to be processed. Content identifiers may be provided during ingestion to enable AMP to export deliverables to the appropriate item in the target system.

Workflows can be set up to meet the needs of the Collection (see discussion on Galaxy, the workflow engine). The AV content can then be submitted to specific workflows, which may include steps for human intervention. The results of the content file processing through the tools added to the selected workflow can be inspected via the AMP Dashboard. Using the AMP Deliverables page, the user selects which outputs to export to the target system (see section below).

Figure 2 shows a screenshot of the AMP dashboard:

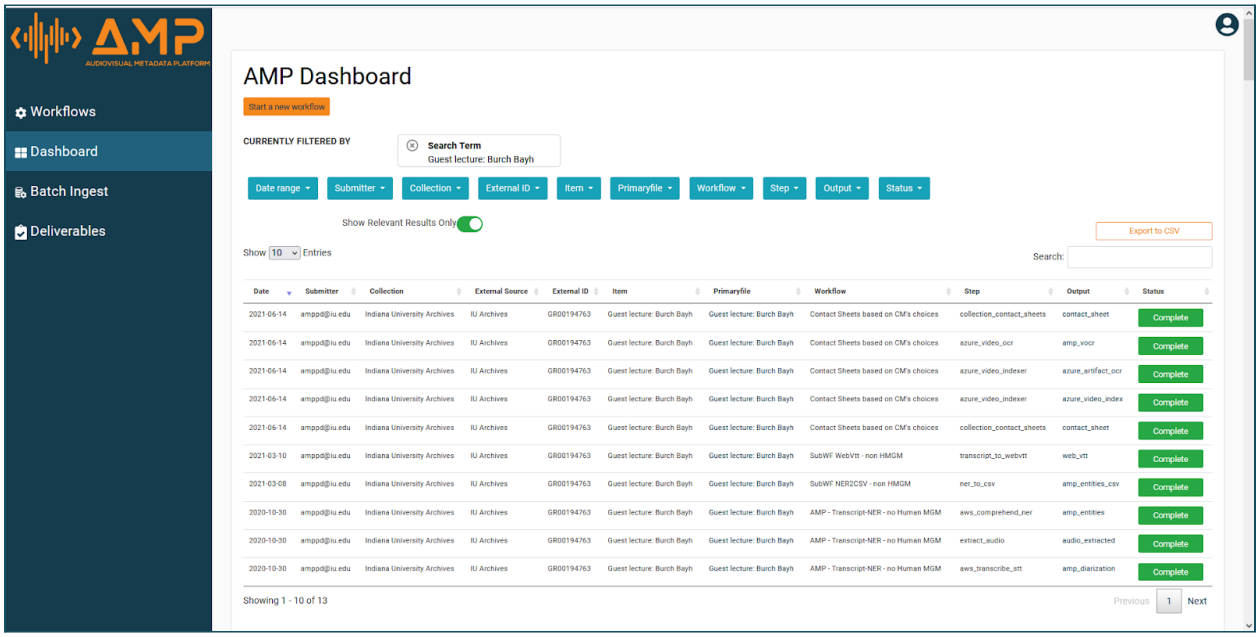


Figure 2. AMP dashboard

Galaxy Workflow Engine

After researching options in the workflow engine space, the AMP development team settled on using Galaxy. The Galaxy open-source, web-based workflow engine is a robust application that provides all the features one expects of workflow engines; among them are the abilities to:

- create, update, and delete workflows

[22] "AMP: Audiovisual Metadata Platform," GitHubGithub, accessed June 22, 2021, https://github.com/AudiovisualMetadataPlatform.

- add new custom tools to be used in workflow steps
- run workflows and resume paused workflows
- tell when a job was run, how long it took, which input files were used, and which parameters were used.

In addition, the engine also validates input file types at time of workflow creation, preventing malformed workflows. Galaxy also manages all AMP job queueing needs.

Figure 3 is a screenshot of the Galaxy interface with a workflow for speech recognition and creation of caption files:

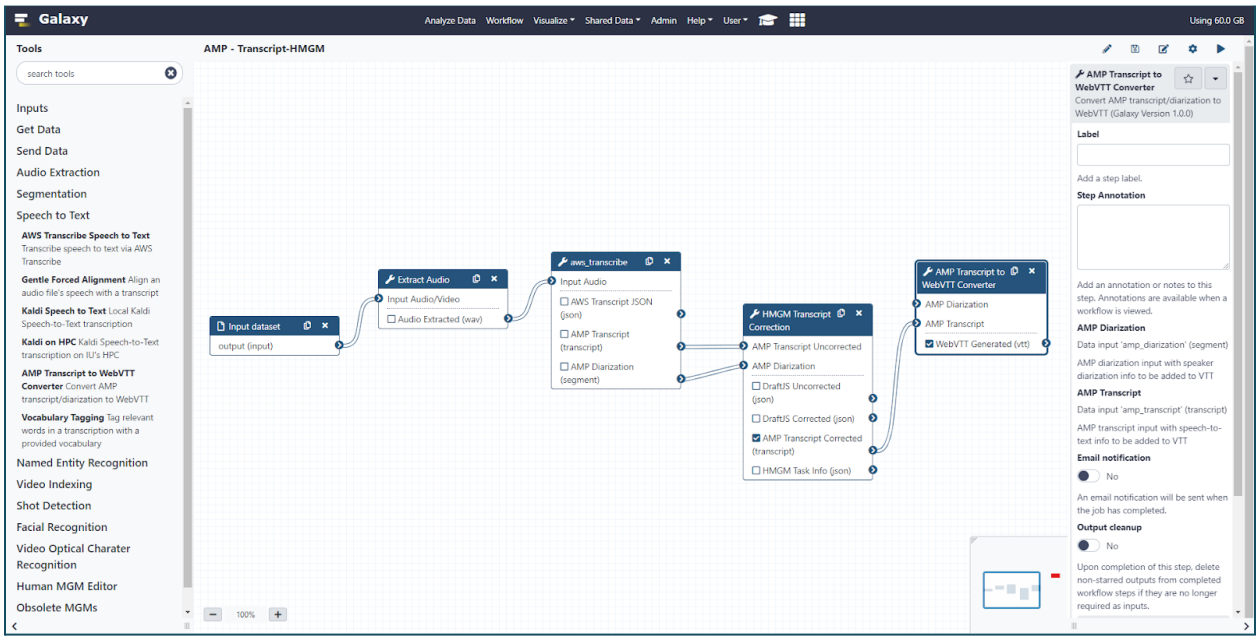


Figure 3. Galaxy interface

Human MGMs

The development team faced technical challenges in the implementation of the human MGMs (HMGMs). Most significantly, Galaxy does not offer out-of-the-box methods to set up workflow steps that may have a long wait period, sometimes several weeks. As delivered, Galaxy locks computing resources until the step is completed in success or error. The team creatively devised a way to use Galaxy's ability to deal with job runners to handle the HMGMs—and the inevitable delays in completing those steps—to resolve the locked resources issue. Once an HMGM job is kicked off, its status is checked periodically at a reasonable interval[23] to see if the

[23] "Reasonable interval" is the time it takes from when users can see HMGM job status updates balanced against how many resources the job runner consumes. Having a shorter interval allows users to see updates sooner, but it causes the job runner to run more often and consumes more CPU and other resources. For example, the current interval is set at 5 seconds per HMGM job, so with 30 HMGM jobs in the queue, this means about 2.5 min of waiting time. For example, if an HMGM transcriber completes an

job has been completed. If it is not done, the job is requeued to allow for further waiting until the task is complete. This frees up computing resources for other MGMs while human tasks are in progress. Additionally, a number of job runners were assigned to balance the overall throughput between HMGMs and automated MGMs.

The second challenge was to integrate the HMGMs with a task management tool. AMP takes advantage of APIs to open and close tickets—if the workflow includes human intervention steps. Once those steps are reached in the workflow processing, AMP creates a ticket in the task management tool for HMGM staff working with the revision tools, which at this point in the project includes the BBC Transcript Editor for transcript correction, and an AMP version of the Avalon Timeliner used to review NER results. The workflow in Galaxy is then placed in “waiting” mode until the human step is completed. Figure 4 shows a screenshot of the BBC Transcript Editor integrated into AMP:

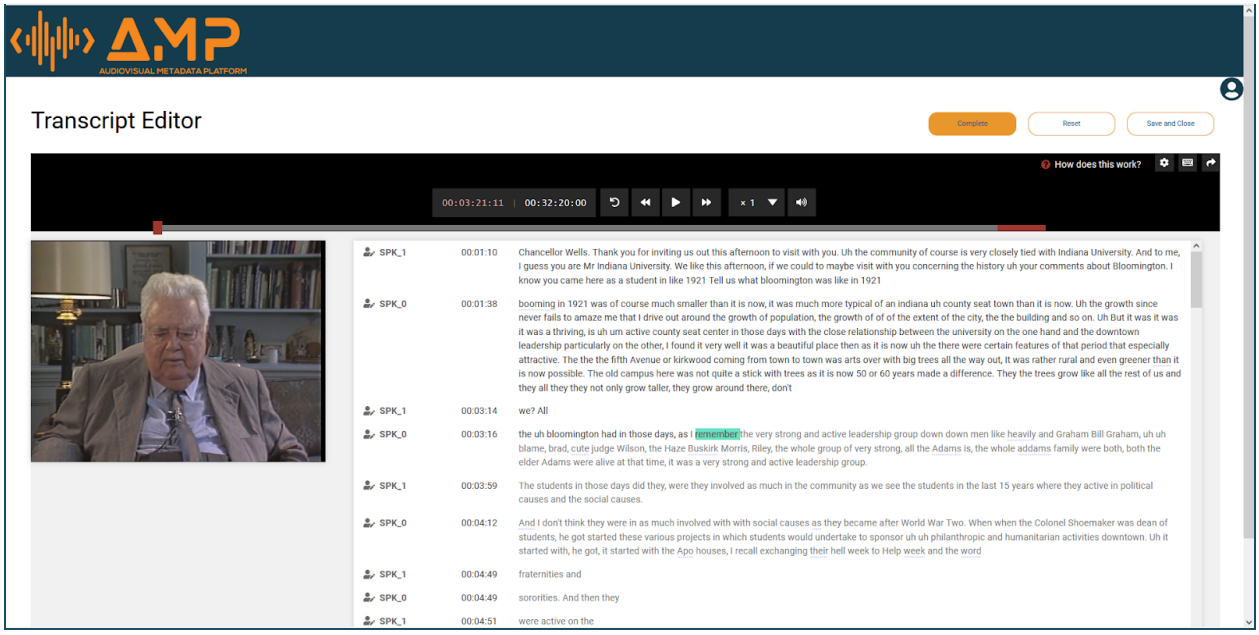


Figure 4. BBC Transcript Editor integrated into AMP

Once staff completes the revision, the AMP application closes the ticket in the task management tool and completes the paused Galaxy step, enabling the workflow to proceed to the next step. Besides the APIs, the integration with task management tools is done by adding item information and links to the task description of the tickets. Staff performing the human intervention activities can use the features native to the task management tool to manage the queue, assign work, and evaluate effort.

HMGM task, she might need to wait for 2-3 min before the Jira ticket is closed. Even so, the interval might still be too short and wasteful from the system point of view, considering that it might take weeks for an HMGM task to complete, during which time the job runner keeps checking while no updates happen.

By the end of the AMPPD project, the only task management tool integrated with AMP is JIRA, but there are plans to integrate with Trello as well since this tool has a free version with APIs and sufficient functionality to meet AMP needs.

Multiple Machine-learning Tool Platforms

One of our goals for AMPPD was to investigate the use of machine-learning and other tools in a variety of platforms: locally, in the cloud, and in high-performance computing environments. We have successfully accomplished this as follows:

Local tools

- spaCy
- Tesseract
- PySceneDetect
- Kaldi STT
- INA Speech Segmenter
- Dlib face recognition
- Gentle Forced Alignment
- Contact Sheet generation
- Multiple file format conversions
- Ffmpeg
- Applause Detection
- Vocabulary Tagging

HMGM tools

- Transcript Editor
- NER Editor

Cloud computing

- AWS Comprehend
- AWS Transcribe
- Azure Video Indexer
- Azure Video OCR
- Azure Shot Detection Generator

High-performance computing (HPC)

- GPU-based Kaldi STT
- GPU-based INA Speech Segmenter

The experience gained with the successful implementation of tools in IU's HPC environment,[24] in particular, allowed the development team to consider enabling other GPU-accelerated options; for example, we have begun discussions for setting up a version of the Gentle Forced Alignment tool in the HPC environment given its resource-intensive profile.

Exporting to external systems

As seen in the AMP architecture diagram, the expectation is that AMP deliverables reach external system(s). Metadata generated in the platform is saved in a database and eventually gets exported to the external access systems that will utilize it. For this phase of the project, the external system being targeted to consume the metadata is the Avalon Media System.[25]

The AMP deliverables are prepared in bags that follow this structure:

```
public class BagContent {
  private Long resultId;     // the id in WorkflowResult
  private String submitter;
  private Date dateCreated;  // job start time
  private Date dateUpdated;  // job end time
  private String workflowId;
  private String invocationId;
  private String stepId;
  private String outputId;
  private String workflowName;
  private String workflowStep;
  private String toolInfo;
  private String outputName;
  private String outputType;
  private String outputUrl;
}
```

A number of APIs are offered to retrieve bags for individual media files, bibliographic items (which may contain multiple media files), or whole collections. AMP also offers the option of pushing the data from the AMP Deliverables page to the external target system. Figure 5 shows the AMP deliverables page:

[24] HPC environments utilize high-end servers and specialized computing hardware (such as GPUs) to provide high-throughput computational resources that can be used for machine learning, big data processing, and other processes that require more computing power than is available on typical servers and workstations.
[25] https://www.avalonmediasystem.org/

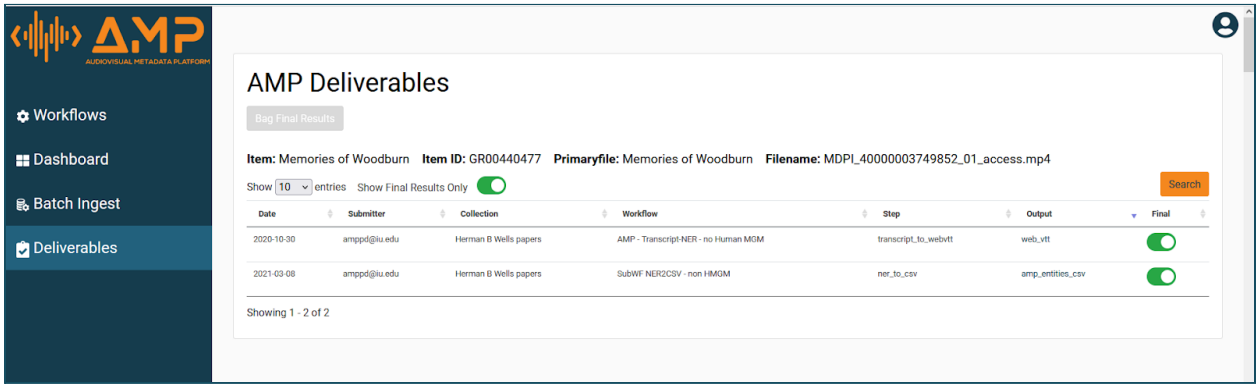


Figure 5. AMP deliverables page

COLLECTIONS

Two collection partners from Indiana University—IU Archives and Cook Music Library—and the New York Public Library (NYPL) participated as an external collection partner. Each participating collection selected 100 hours of audio and video content. IU Archives focused on scholarly presentations and lectures, along with recordings of historical figures and events within the university. Cook Music Library focused on various musical performances like jazz ensembles, operas, solo recitals, and orchestra performances. NYPL selected items specifically from the Gay Men's Health Crisis collection[26] that included protests, speaker events, interviews, and focus groups. Each collection's video content took up to 75-80 hours of their 100 hours of content, with audio-only content making up the remaining 20-25 hours. The goal was to provide the widest variety of speaking combinations (formal/informal, single/multiple speakers, orderly/disorderly conversations, clear audio/less clear audio), musical types (single instrument, multiple instruments, singing, performances with applause and speaking in between, varying tempos and musical styles), and video settings and styles (professionally edited, amateur footage, stage performances with actors in costume, individual people in frame, multiple people in frame) to represent the collections being used and offer options for trying different types and qualities of audio and video against various MGMs for thorough analysis. Each collection also had metadata goals that factored into the selected 100 hours of content: IU Archives provided a set of video items that included Herman B Wells, a major figure in Indiana University history, so they could determine how searching for a known person might work; Cook Music Library wanted to figure out easier ways to segment audio items based on musical segments; and NYPL had an interest in easily detecting all contents on single video item.

After items were selected for each collection, the collections staff worked with the MGM team to narrow down the types of MGMs that might produce the desired metadata outputs. As categories of MGMs were determined, collections staff helped to review results from different tools in each category (comparing transcript outputs and named entity recognition outputs, for example). After MGM tool decisions were made and those MGMs were incorporated into the

[26] "AIDS Activist Videotape Collection," New York Public Library, accessed June 28, 2021, http://archives.nypl.org/mss/3622.

AMP system, collections staff participated in evaluating both the AMP interface and the outputs for their collection content to determine how well the AMP system worked for their needs and if the MGM workflows were able to produce useful metadata to enhance their collection description.

FINDINGS

MGMs

Challenges with Proprietary Tools

The black-box nature of many proprietary tools was a significant challenge. For example, not knowing what, if any, preprocessing was happening in tools like AWS Transcribe and Google Cloud Speech-to-Text made it difficult to test whether additional preprocessing of audio would improve results. The team also encountered unhelpful error messages (or a lack of error messages altogether) that made problems unresolvable. For instance, Azure Video Indexer would often fail to process black and white videos with no feedback as to why, and Google's speech-to-text and diarization tool would consistently fail on some files with no error message.

One particularly concerning black-box process required significant alteration to how outputs were handled. The development team discovered late in testing that AWS Transcribe speech-to-text transcription was censoring what it deemed "offensive" words in transcripts without opting in to that feature and without providing the list of censored words. While this practice may serve the use cases of some AWS customers well, truthful transcription of content is imperative in the world of archives. All collections staff agreed that they would prefer to have potentially harmful words transcribed verbatim but flagged for their attention, so they could address their visibility downstream with trigger warnings or other mechanisms for alerting patrons and staff to their presence. After several rounds of email communication with numerous levels of AWS customer support (which required upgrading to an enterprise plan), the project team learned that AWS keeps an internal list of words to replace with asterisks. Customers may opt out of this practice, but it was unclear how they would know where to make this request, especially those customers not paying for a plan that allows direct email communication with customer support.

Though a disappointing experience with proprietary tools, this incident did draw the project team's attention to a risk not previously considered—how to flag potentially harmful content that is generated by an MGM (either truthfully reflective of the content or created in error) before sending it along the pipeline to human MGMs or public users. This inspired the project team to design a new MGM—a simple vocabulary filter that takes a list of words from a collection manager and flags them in an MGM text output. This MGM can be used to identify not only potentially harmful words but also words of interest that support other use cases.

Challenges with Determining Accuracy

For some MGMs, success depended upon how well the results satisfied a specific use case. Different collections had different use cases that required different levels of accuracy and output formats. For example, one collections staff person wanted the shot detection MGM to detect different pieces of content recorded on one physical tape or disc; another wanted a visual summary of the contents as a quick reference for access to the file or as a proxy in case of damage to the video file. The former application required a higher level of accuracy at a lower granularity (fewer frames analyzed), looking for only a change in content. The latter did not require high accuracy but did require higher granularity, showing representative frames from all shots within the same recorded event rather than only major changes of content. This exercise underscored the need to generate ground truth and to test for accuracy based on use case rather than trust any one metric to predict the usefulness of a tool.

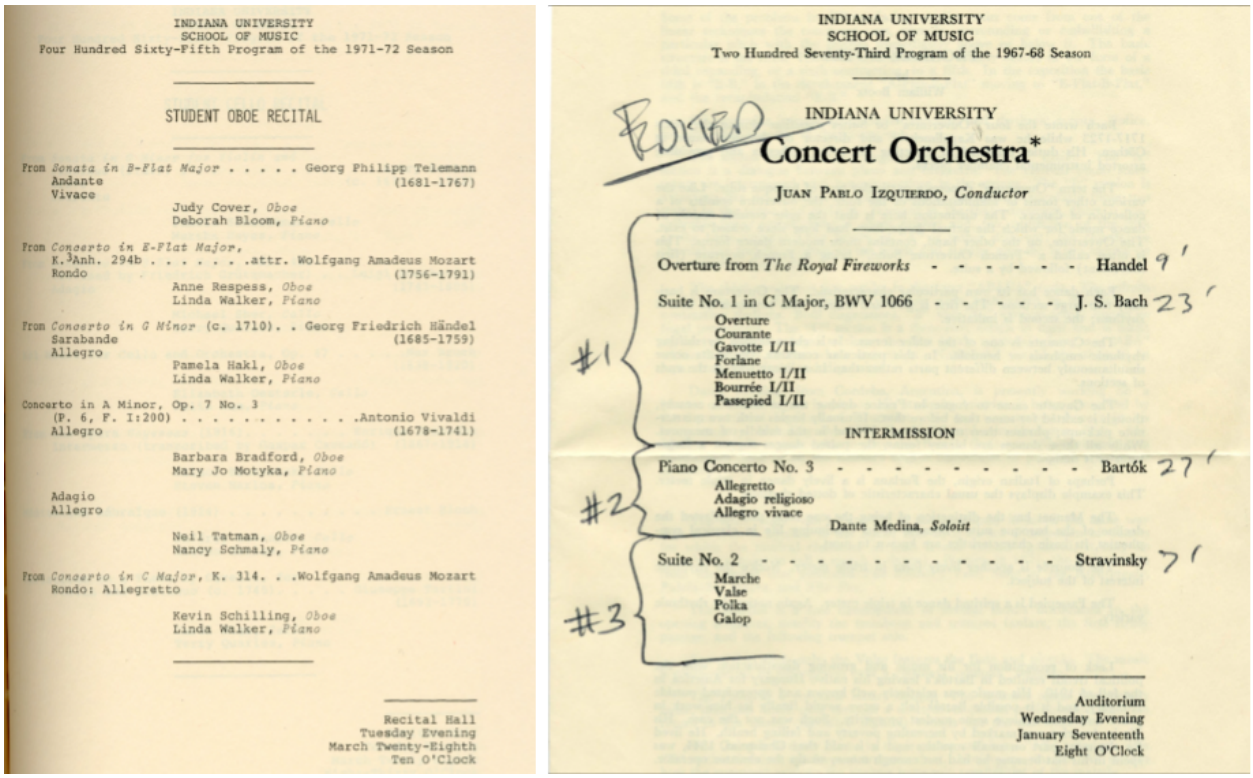
Challenges with Less Common Tasks

Not all MGM ideas generated at the start of the project were implementable or practicable enough to meet the needs of the Collection team's use cases. Music genre classification had the problem of being highly subjective and difficult to integrate with traditional cataloging practices. Tools have different concepts of genre, with different levels of granularity. Similarly, vocabularies used to describe genre vary by collection and descriptive practice. The granularity of genre used by a general music library collection may be different from that used by a culturally-specific music collection. As evidenced by the wide range of music genre terms appearing in the Library of Congress Subject Headings, the level of granularity of genres varies widely. Additionally, many genre terms reference specific nationalities, a distinction at a level of granularity that may be quite difficult for an algorithm to make. While large companies like Spotify may have the resources and corpus to be able to classify genre at a fine level of granularity, openly available models or training samples for classifying genres at a level of granularity to be useful to music catalogers are still very difficult to find. As digitized, cataloged holdings of libraries become more publicly available and easily downloadable, the potential for building models for genre classification may become more real, though the taxonomy of LCSH music genre terms[27] applied may need a complete rethinking to be more easily distinguishable by a classification algorithm.

Running zonal optical character recognition (OCR) on printed music programs to extract structured information on performers and works performed presented a similar problem of differences across collections. Most tools that purport to extract structured data from documents via OCR are made for standardized document layouts with key-value pairs in form fields. The MGM team attempted to extract performer names listed in blocks below their instrument and to identify works and their composers, separated on the same line by a series of dots. While these layouts appeared to the human eye to be fairly consistent, there was enough variation across

[27] "Genre/Form Terms for Musical Works and Medium of Performance Thesaurus," Library of Congress, accessed June 28, 2021, https://www.loc.gov/catdir/cpso/genremusic.html.

programs to confuse the algorithm and produce inconsistent results. While the team achieved limited success on this task, they only tested one tool, Tesseract,[28] due to a lack of open-source or reasonably-priced proprietary solutions. More work could be done to explore tools like ABBYY FineReader,[29] which was prohibitive in its cost for this brief experiment but is known for its advances in this area.



INDIANA UNIVERSITY
SCHOOL OF MUSIC
Four Hundred Sixty-Fifth Program of the 1971-72 Season

STUDENT OBOE RECITAL

From *Sonata in B-Flat Major* . . . . . Georg Philipp Telemann
Andante (1681-1767)
Vivace
Judy Cover, *Oboe*
Deborah Bloom, *Piano*

From *Concerto in E-Flat Major*,
K.[3]Anh. 294b . . . . . .attr. Wolfgang Amadeus Mozart
Rondo (1756-1791)
Anne Respess, *Oboe*
Linda Walker, *Piano*

From *Concerto in G Minor* (c. 1710). . Georg Friedrich Händel
Sarabande (1685-1759)
Allegro
Pamela Hakl, *Oboe*
Linda Walker, *Piano*

Concerto in A Minor, Op. 7 No. 3
(P. 6, F. I:200) . . . . . .Antonio Vivaldi
Allegro (1678-1741)
Barbara Bradford, *Oboe*
Mary Jo Motyka, *Piano*

Adagio
Allegro
Neil Tatman, *Oboe*
Nancy Schmaly, *Piano*

From *Concerto in C Major*, K. 314. . .Wolfgang Amadeus Mozart
Rondo: Allegretto
Kevin Schilling, *Oboe*
Linda Walker, *Piano*

Recital Hall
Tuesday Evening
March Twenty-Eighth
Ten O'Clock

INDIANA UNIVERSITY
SCHOOL OF MUSIC
Two Hundred Seventy-Third Program of the 1967-68 Season

INDIANA UNIVERSITY
Concert Orchestra*
JUAN PABLO IZQUIERDO, *Conductor*

#1
Overture from *The Royal Fireworks* - - - - - - Handel 9'
Suite No. 1 in C Major, BWV 1066 - - - - - J. S. Bach 23'
Overture
Courante
Gavotte I/II
Forlane
Menuetto I/II
Bourrée I/II
Passepied I/II

INTERMISSION

#2
Piano Concerto No. 3 - - - - - - - - - - Bartók 27'
Allegretto
Adagio religioso
Allegro vivace
Dante Medina, *Soloist*

#3
Suite No. 2 - - - - - - - - - - - - - Stravinsky 7'
Marche
Valse
Polka
Galop

Auditorium
Wednesday Evening
January Seventeenth
Eight O'Clock

Figure 6. Variations in concert program layout. While visibly similar to the human eye, the layouts were not similar enough in spacing and separating dots/dashes for Tesseract's zonal OCR to produce consistent results.

As might be expected, the team had greater success with more common machine-learning problems, such as speech-to-text transcription and named entity recognition, where commercial tools have been developed around extensive research. More specific problems that are less marketable, like music genre classification or applause detection, usually are not addressed by a commercial solution, so the only tools available are often small open-source projects or codebases shared as academic research outputs. The project team worked with the Fraunhofer Institute for Digital Media Technology[30] to try to fill some of these gaps in music analysis with some of their commercial tools, but the level of granularity of classification offered was too broad to be of use to music catalogers.

[28] "Tesseract," AMP wiki, accessed June 28, 2021, https://wiki.dlib.indiana.edu/display/AMP/Tesseract.
[29] Homepage, ABBY Fine Reader, accessed June 28, 2021, https://www.abbyy.com/ocr-sdk/.
[30] Homepage, Fraunhofer Institute for Digital Media Technology, accessed June 28, 2021, https://www.idmt.fraunhofer.de/en.html.

Successes with Machine Learning

While training a model was not a goal of this project originally, for some of these niche problems this proved to be the only solution. The MGM team successfully trained custom models for two categories: applause detection and facial recognition.

Developing a classification algorithm for applause grew out of a Music Library use case for identifying index points for works in an audio or video performance to aid users in navigating within the content. While the music, speech, and silence detection MGM could separate segments of music from non-music, this did not help distinguish musical works, which may be composed of multiple discrete segments of music separated by silence. In many types of musical performance, the end of a work is followed by applause, so it was hypothesized that detecting the applause within a work could help a human user set these index points more efficiently (with the understanding that results may include some false positives—applause when performers take the stage or applause after jazz solos, for example). After experimenting to little success with YAMNet,[31] an existing model for detection of over 500 sounds, including applause, as well as a model the project team contracted the Fraunhofer Institute to train, the MGM team decided to train their own model, adapting the Acoustic Classification & Segmentation model developed by the Brandeis Lab for Linguistics & Computation[32] to classify audio segments not only as speech or non-speech, but speech, music, silence, noise, and applause. The model was trained using three-second samples from the MUSAN corpus[33] of music, speech, and noise; three-second applause samples from the HIPSTAS project;[34] and three-second samples of all five categories from our partner collections' files. The resulting model was very effective at classifying applause, but less effective at distinguishing speech, music, noise, and silence. The MGM team modified the code to create a binary applause/non-applause classifier that could be used as a separate MGM to supplement other segmentation tools.

A similar process gave rise to the development of a facial recognition model. Collections staff showed interest in the ability to locate specific known individuals in their collection materials. While commercial solutions for facial recognition exist, the project team was wary of the ethical implications of handing a corporate entity face recognition data on people without knowing how that data might be used in the future. The MGM team identified an open-source Python library[35] for facial recognition to create an MGM that allows AMP users to supply their own images of the

---

[31] "Yamnet," Github, accessed June 24, 2021, https://github.com/tensorflow/models/tree/master/research/audioset/yamnet.
[32] Brandeis University Lab for Linguistics and Computation, "Acoustic Classification & Segmentation," Github, accessed June 24, 2021, https://github.com/brandeis-llc/acoustic-classification-segmentation.
[33] "MUSAN," Open Speech and Language Resources, accessed June 24, 2021, https://www.openslr.org/17/.
[34] HIPSTAS, "Applause Classifier," Github, accessed June 24, 2021, https://github.com/hipstas/applause-classifier.
[35] "face-recognition 1.3.0," Python Package Index (PyPI), accessed June 24, 2021, https://pypi.org/project/face-recognition/.

person they want to find in videos and train a model specifically for that individual. Although these models are only as good as the data given them, this workflow lowers the barrier of entry for content experts to be involved in training machine-learning models.

While more advanced knowledge of machine learning is necessary to tune models for optimum performance, the success of the MGM team in producing reasonably accurate results with limited experience in model training points to an area of potential for empowering collections staff to train models within AMP for specific purposes. With a wealth of potential training materials within archives, sufficient guidance on selecting appropriate samples for training, and simplified mechanisms for creating and testing against ground truth, collections staff could become active participants in designing new MGMs that meet some of the more niche needs for processing and description in archives.

Successes with Process

The collaborative process used by the MGM team over the course of the project proved effective, in large part due to the inclusion of a liaison to the collections staff, the product owner, the project manager, and additional SMEs in weekly discussions about individual MGM evaluations.

Human MGMs (HMGMs)

Efficiency

There are a number of factors to consider when trying to evaluate the overall efficiency and impact of supplementing automated MGM workflows with HMGMs for transcript correction and named entity revision. One area to consider is the amount of time it takes for transcript correction. Some professional transcription services estimate that, on average, it takes four hours to transcribe one hour of content[36]. However, this estimate comes with some rather large caveats regarding the quality of the audio, the number of speakers, the accent of the speaker(s), and background noise. Additionally, the transcription process can be slowed if extensive research is required on behalf of the transcriber to accurately identify speaker identities and unfamiliar terms.

In AMPPD, the HMGM staff performed transcript correction at an overall rate of approximately 10.5 hours per one hour of content (~159 correction hours for ~15 hours of content). While it is possible that starting with the automated speech-to-text transcript and correcting rather than creating a transcript from scratch was generally an efficient strategy, it was not enough to counteract some of the hurdles faced in AMPPD's HMGM workflow, as described below.

[36] "How Long Does It Take to Transcribe One Hour of Audio or Video?," Rev.com, accessed June 24, 2021, https://www.rev.com/blog/resources/how-long-does-it-take-to-transcribe-audio-video.

Staffing

The team chose to use IU cataloging staff for our HMGM work. While familiar and comfortable with detailed work, our HMGM staff were not overly familiar with transcription. There was a learning curve with transcription as an activity, as well as with our selected tools that certainly contributed to an increased amount of time spent on transcript correction at the beginning of the project. Furthermore, HMGM staff did not have the same expertise with the nature of the content from the partner collections as collections staff might. Not having familiarity with the collections and the people, places, topics that might be included as subjects led to increased time and effort on the part of editors, who needed to conduct external research (e.g., to look up correct spelling of proper nouns). In the future, it may be more feasible to use existing collections staff to perform transcript correction, especially if they have transcription experience.

Content Selection

The content that was selected to go through the HMGM workflow was not always the most appropriate. By selecting a semi-random sample of content based on length of recording, there happened to include some items that were not the best suited for human correction. This includes recordings with poor audio quality, where automated speech-to-text may be just as accurate, and raw footage of events that would not necessarily benefit from a transcript. These items were passed through the transcript correction HMGM with minimal intervention. The efficiency and impact of HMGMs can be improved by increased selectivity in the content chosen to receive human intervention.

Tools

The efficiency of any workflow is highly dependent upon the performance of its supporting tools. In the case of HMGM workflows within AMPPD, the BBC Transcript Editor proved to be a significant obstacle. HMGM staff reported several bugs with the editor during the course of the pilot, including significant timestamp synchronization problems, cursor and scroll bar oddities, and crashes due to lack of memory. Bug reporting and bug fixing with the editor was happening at the same time as transcript correction was taking place. This led to frequent interruptions in HMGM work and increased the amount of time spent correcting each transcript, especially at the beginning of the project. Human-corrected transcripts also tended to include a wider variety of punctuation marks than the AMP code to transfer corrected transcripts to natural language processing tools anticipated. This resulted in a delayed start to the named entity revision HMGM. Future improvements to the performance of integrated HMGM tools, particularly the BBC Transcript Editor, will greatly increase the efficiency of these workflows.

ETHICAL CONSIDERATIONS

Ethical and privacy concerns were always in front of mind for the project team as they analyzed MGM tools. They have always had the concern of introducing bias via the machine learning algorithms and training data used by the tools, particularly commercial tools that are often not

transparent about their training data sets or how those data sets are used to produce outputs. Learning from resources such as Dr. Safiya Noble's *Algorithms of Oppression* and Joy Buolamwini's work and research with the Algorithmic Justice League regarding the problems of artificial intelligence algorithms showing bias based on skin color, gender, and race, particularly from commercial vendors, reinforced our sense of discomfort in supplying any content to these services for facial recognition or speaker identification and trusting in the results they produce.[37] This directly resulted in an adjustment in scope for facial recognition, as well as close inspection and questioning of results from automated transcription and named entity recognition tools.

After learning about the problems of proprietary facial recognition data sets and surveillance and privacy concerns, preference was given to the open-source Python-based tool face_recognition.[38] Its inner workings are open and it only recognizes faces that are based on supplied training images of a known person, as opposed to trying to identify any face that appears within an unknown or proprietary data set. This fits the IU Archives' use case in which a set of videos can be searched for a known person (Herman B Wells) instead of trying to identify unknown faces using an unknown or proprietary data set from, for example, Amazon Web Services.

Ethical considerations also highlighted the essential role filled by people in using AMP's Human MGM tools for evaluating and correcting automated outputs from various automated MGMs. It was through human intervention that one automated transcription tool (Amazon Transcribe) was found to be removing "harmful" language without our knowledge. The resulting discussion and investigation, which involved reaching out directly to Amazon's technical support team (no easy task), found that the collections staff did not want to remove potentially offending words from the transcript due to the need to acknowledge the reality of history and what was recorded. Additionally, removal of the terms meant they would not be discoverable. Keeping human review steps in the AMP workflow helped ensure our understanding of automated MGM outputs were used and what they actually did.

TECHNOLOGY

There were several choices the development team had to make as they designed and developed the system. Choosing Galaxy as the workflow engine has proven to be a strength of the project. Because of Galaxy's features, the team has been able to implement every tool selected without issues and integrate it with JIRA and with IU's High Performance Computing (HPC) environment. Additionally, Galaxy offers numerous APIs that make integration with the AMP-specific code possible, and it offers ways to extend its functionality, which enabled the implementation of the HMGMs. Another important Galaxy feature is its ability to do file-type checking during workflow design to prevent the creation of malformed workflows. This is an excellent feature, but the file types native to Galaxy are not refined enough for AMP purposes.

[37] Homepage, Algorithmic Justice League, accessed June 23, 2021, https://www.ajl.org/.
[38] "Face Recognition," Face Recognition documentation, accessed June 23, 2021, https://face-recognition.readthedocs.io/en/latest/readme.html.

For instance, Galaxy considers any binary file just a binary file, whether it is media, document, or another file type. To illustrate the problem, AMP provides tools that require audio as input and tools that require video as input; with the native file types, one would be subject to runtime errors due to file type mismatch. On the other hand, Galaxy also provides the ability to define new file types, and AMP is taking full advantage of this feature to address not only the media scenario discussed above, but also to refine the JSON types. All AMP tools generate outputs in JSON format, but the schemas may differ; e.g., a JSON schema representing a transcript output is different from the JSON schema representing extracted named entities. Being able to refine data file types is essential to AMP and allows the application to take full advantage of Galaxy's type checking during workflow creation to provide the best user experience possible.

Another happy decision was integrating with external components rather than recreating them internally. The system architecture outlined in the platform architecture workshop conducted in 2017-2018, during the project's planning phase, differs from the current architecture precisely in which components we chose *not* to write. We accomplished this in two ways:

1. Via APIs—this is how AMP is integrated with the workflow engine component (Galaxy) and the task management component (JIRA).
2. Via packaging—this is how we integrate the tools for human review of automatically generated output, i.e. the BBC Transcript Editor and the Avalon Timeliner.

During the course of the project, the development team had to make choices about where to focus our efforts. When obliged to choose between 1) improving and adding to the AMP user interface, or 2) implementing features in AMP or Galaxy to further prove the concept of a platform for metadata generation that allows for human intervention steps, we chose the latter, using backdoor options to make up for the lack of user interface features. This left us with a reasonable amount of frontend development to do in the next phase. As an additional consequence, collections staff were unable to fully experiment with the system, depending on the Scrum team for workflow submission of their content.

Focusing on proving the concept also led us to create some technical debt. For instance, our load testing efforts to date have brought up issues that we have not had the chance to address yet. Preliminary discussions have pointed to possible solutions to explore in the next phase.

The integration with the BBC Transcript Editor[39] has been helpful, but it is also a challenge. It was chosen for its simplicity and open source nature. However, as of publication, it is not actively maintained and, given that the code was a prototype for the BBC, it is not altogether well structured or easy to change or fix. One significant problem is its inability to adjust the transcript timestamps when more than a few words of text have to be added as part of the correction process. The solutions we found were all workarounds. When looking for a transcript editor to use, we did not find another option, but the experience with this project does underline

[39] "BBC React Transcript Editor," Github, accessed June 23, 2021. https://github.com/bbc/react-transcript-editor.

the importance of an active community of maintainers when it comes to open-source code (which is one of the great benefits of Galaxy).

OUTPUTS

Output formats from the audiovisual content that has been processed by the AMP system are varied depending on the MGM(s) and needs defined by the collections staff, which were gathered throughout the project. Outputs range from CSV files, to JSON, WebVTT, and contact sheets (screen shots grabbed at defined intervals delivered side by side in a single PNG file). Outputs were delivered to collections staff throughout the second half of the project to ask for feedback, and were adjusted based on their needs and preferences. This section walks through the development of the outputs as they relate to these preferences.

Collections staff were asked to review the outputs from their 100 hours (each) of audiovisual content beginning in late 2020. The outputs included contact sheets of video content, video OCR, transcripts for both video and audio-only content, and named entity recognition outputs based on those transcripts. Feedback was gathered initially via a survey form, followed by an online feedback workshop examining sample data outputs from each collection, and finally through conversations and email. The survey results indicated that all three collections representatives were able to locate and accurately identify workflows and outputs in the AMP interface (from which the outputs are delivered), however one facet of the site—the AMP Dashboard—received mixed ratings. Specific requests from collections staff included:

- search filters need to be *facets that are browsable* (they are only searchable)
- there are accessibility issues with color contrast, link styling, and results paging functionality
- there were too many named entity recognition file outputs available (CSV and JSON output files were both showing at the time of the survey).

Most of these issues have since been resolved; replacing some of the filters with a facet interface will take place as part of the UI work in the next phase of the project.

Named Entity Recognition (NER)

The collections staff found most of the data outputs extremely useful, with the exception of named entity recognition, which initially were presented in JSON. Collection staff found the JSON format difficult to navigate. As with other MGM outputs, the development and MGM teams worked iteratively to respond to collections team feedback. In this case, an alternative output format was offered to them—a CSV file. Collections staff found the CSV file to be much more user friendly, although this format highlighted some out-of-the-box quirks. The main concern was the categorization of terms, which in some cases were not useful. For example, in all cases, collections staff found the category “QUANTITY” to be more or less useless, and “DATE” to be unpredictable. For example, one collection staff person commented that “Some categories produce less useful results, e.g., ‘DATE: today’.” Additionally, values in the QUANTITY category included “18” and “94”—out of context these values are meaningless.

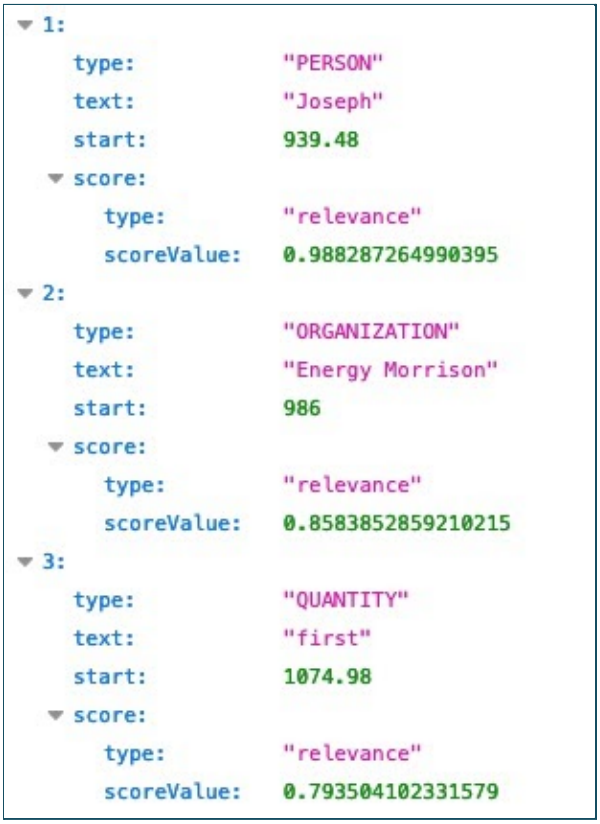


Figure 7. Excerpt from NER JSON file showing the categories PERSON, ORGANIZATION, and QUANTITY.

Collections staff also mentioned that the CSV file "[d]oes not include [the number] of occurrences of terms, which would be helpful in determining relevancy." When they realized they could sort the CSV by category and by term, collections staff responded positively, suggesting several ways the output could be used: to "[i]dentify potential sources for researchers," for "guidance for more in-depth description," and for "quick subject tags." Future enhancements to this MGM include giving the collections staff the option to include or exclude particular categories to better hone in on the data about which they are particularly interested. This functionality is already available in AMP using either AWS Comprehend or spaCy[40]. In the next phase of the project we will make sure users understand how to utilize it.

DATE	ago	1921.98
EVENT	Thanksgiving	617.95
LOCATION	Madison, Wisconsin	140.37

Figure 8. Excerpt from NER CSV file showing the categories DATE, EVENT, and LOCATION, and associated terms. The numerical value is the point in the audio where this instance of the term appears.

[40] "MGM — Entity Extraction," AMPPD wiki, accessed June 25, 2021, https://wiki.dlib.indiana.edu/display/AMP/MGM+-+Entity+Extraction.

Transcription (Speech to Text)

Transcriptions were produced using AWS Transcribe and Kaldi[41]. As with NER, the original JSON outputs were extremely difficult for collections staff to use directly, and initial reviews were not positive. An additional challenge was the general quality of the data produced by the MGMs. After discussions, the development team, as they did with NER, provided the output in a second, more user-friendly format: WebVTT.[42] WebVTT is a standard for captioning video, presenting extracted text with timestamps in a simple text file, which makes it extremely clear where the text occurs in the video file. Once the data was provided in this format, collections staff found it much easier to navigate. One collections staff wrote, "The WebVTT format is pretty easy to understand/navigate, and I think it [is] sufficient in terms of a transcript for internal use. I know that all of the time spans are so short/frequent because it is formatted for captions, but one thing that would make it more readable as a transcript would be if there were less frequent time indicators." This suggestion is a good one, and definitely achievable. The development team will consider implementing this option in the next phase of the project.

It should be mentioned that the HMGM team did a great deal of clean-up on some of the speech-to-text outputs. Unsurprisingly, once they had corrected the transcripts, collections staff were even more pleased by the results.

```
WEBVTT

00:00:43 --> 00:00:47
<v spk_0> Indiana University Television, in cooperation with the Poynter Center,

00:00:47 --> 00:00:57
<v spk_0> presents Conversations on America, a series of programs examining American

00:00:57 --> 00:01:05
<v spk_1> institutions. On this program are David Broder,

00:01:05 --> 00:01:07
<v spk_1> associate editor for the Washington Post,

00:01:07 --> 00:01:10
<v spk_1> a Pulitzer Prize winner who has been described as America's
```

Figure 9. Excerpt from WebVTT file.

Facial Recognition

In general, collections staff members were hesitant to take advantage of this MGM due to the ethical concerns outlined in other sections of this document. However, when the approach was

[41] "MGM — Speech-to-text," AMPPD wiki, accessed June 25, 2021. https://wiki.dlib.indiana.edu/display/AMP/MGM+-+Speech-to-text.
[42] "WebVTT: The Web Video Text Tracks Format," W3C, accessed June 25, 2021, https://www.w3.org/TR/webvtt1/.

adjusted wherein the tool (Python face_recognition[43]) was trained on a single individual, the AMPPD team, in general, felt more at ease. The MGM was successful in identifying one individual (Herman B Wells) in multiple videos, at different ages and different views (facing forward, in profile, seated/standing, etc.). In general, the staff was pleased with the outputs.

Figure 10. Excerpt from Herman B Wells facial recognition contact sheet.

Scene detection

The outputs for scene detection were initially viewed by collections staff as too large to actually see and understand. They were, in most cases, “overwhelmed” by the results. This was mainly due to the frequency of capture; every scene change produced one new capture. For example, an hour-long video could include hundreds of scene changes and produce as many captures. The development team used this feedback to adjust the parameters of the MGM to reduce the number of images. This was greeted with very positive feedback from the collections staff. Responses included, “[t]hey're exactly what we were hoping for—they're tremendously helpful in reviewing material for content and oddities" and “the contact sheets were a lot more digestible and made it very easy to tell when the content on the media had run out leaving only a blank

[43] “MGM—Facial Recognition,” AMP wiki, accessed June 25, 2021, https://wiki.dlib.indiana.edu/display/AMP/MGM+-+Facial+Recognition.

screen. Having that is useful to easily know where to stop the tracking so staff is not staring at the blank screen until the end 'just in case there's something more'." An interesting improvement to the MGM was suggested by a member of the collections staff and could be taken up in the next phase of development—"If you created a tool that could capture any text to add to the contact sheets, I feel like that might be enough and the video OCR wouldn't be necessary."

Figure 11. Excerpt from scene detection contact sheet.

Video OCR

In general, video OCR quality was poor for the sample files. Most sample files were several decades old, copied from VHS or similarly unstable formats, and while the visual quality of the videos themselves was varied, most were less than ideal for this MGM[44] to perform well. The hope was that video OCR would capture text on protest signs, buildings, and street signs, as

[44] "MGM—Video OCR," AMPPD wiki, accessed June 25, 2021, https://wiki.dlib.indiana.edu/display/AMP/MGM+-+Video+OCR.

well as titles and credits. Unfortunately, the MGM did not perform well in many cases, although titles and credits were more successful than the other forms of visual text. For that reason, the collections staff did not find the output useful and, in fact, one commented that the video OCR output was of “poor quality,” and that it would be “better to generate contact sheets for parts that have text on screen.” It is unlikely that, without a great deal of training, video OCR will be useful at scale for archival video collections of the sort used in this testing.

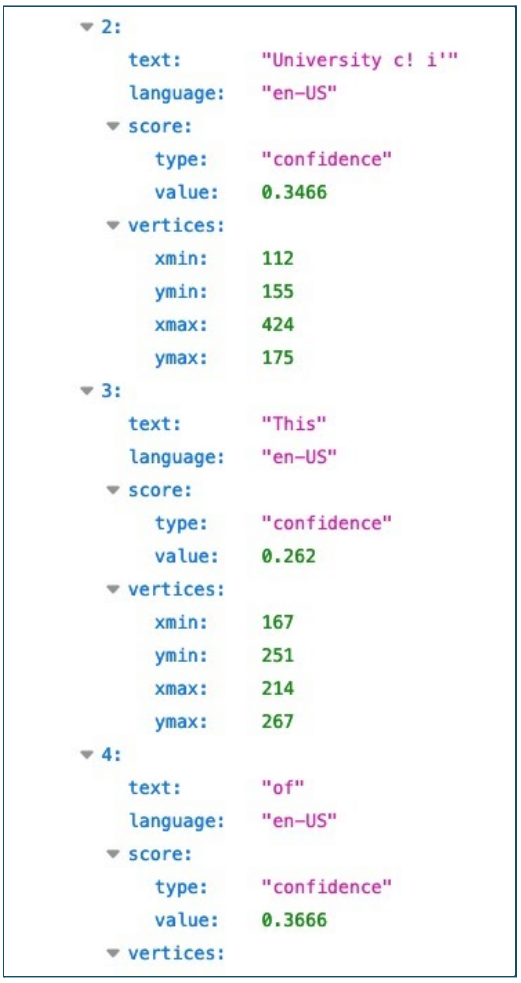


Figure 12. Excerpt from video OCR output. This represents the mid-range quality of output.

At the end of the collections feedback workshop, we posed a “temperature check” question to staff from the four collections used in testing: “[t]hinking about your collections, how do you feel about how we've been able to support your interests since the start of the grant?” Collection managers were asked to record their rating on an image of a thermometer. Staff from three of the four collections scored their experience as “hot” (positive), while staff from the Music Library—not unexpectedly—responded with a lower score. In this phase of the project, the major takeaway from outputs analysis was that, though expectations should have been tempered at the outset, with feedback and iterative development, the team was able to ethically

produce outputs that were useful for the identification of audiovisual content and the enhancement of existing metadata and catalog records in many cases, especially for collections involving significant amounts of spoken-word content.

While collection managers are skilled at developing cataloging and archival processing workflows for humans, workflows for AI will require new types of metadata creation pipelines and new ways of assessing the quality of data outputs to effectively integrate AI into production. In the next phase of AMP, we will develop a module for AMP that empowers and educates collection managers in better evaluating each MGM's suitability for their unique collections and use cases.

CONCLUSION

With these results in mind, we proposed a third phase of AMP that will focus on IU and AVP working together to make the system production-ready for use by IU and other institutions that have needs for describing large quantities of audiovisual content, building a module for collection managers to use to evaluate MGMs for suitability, and using the system to help make additional collections from IU and NYPL more discoverable and usable by researchers. These collections will be selected with a focus on materials from historically underrepresented cultures and populations, keeping in mind the ethical considerations inherent in working with and identifying appropriate access for such collections. In June 2021, the Andrew W. Mellon Foundation generously accepted our proposal, and the next phase of the project began in July 2021. We will continue to report on findings and look forward to feedback on this and future publications.

REFERENCES

AMPPD: Audiovisual Metadata Platform Pilot Development Proposed Activities and Rationale, 2018, https://go.iu.edu/4eSq.

Audiovisual Metadata Platform Pilot Development (AMPPD), https://go.iu.edu/amppd.

Dunn, Jon; Juliet Hardesty. *AMP-lifying IU's media collections with the Audiovisual Metadata Platform*, 2018, http://hdl.handle.net/2022/22564.

Dunn, Jon W.; Juliet L. Hardesty, Tanya Clement, Chris Lacinak, and Amy Rudersdorf. *Audiovisual Metadata Platform (AMP) Planning Project: Progress Report and Next Steps*, 2018, http://hdl.handle.net/2022/21982.

Noble, Safiya Umoja. *Algorithms of Oppression: How Search Engines Reinforce Racism.* New York: NYU Press, 2018.

APPENDIX A. DEVELOPMENT ROADMAP

Milestone #1 - May 7, 2019

Sprints 0-5

1. Dev Tools and infrastructure
2. System Architecture
3. Initial Data Model design
4. File System Layout
5. Basic Ingestion
6. Start conceptualizing the workflow
7. Collections to use
8. Definition of MGMs we will need to address collection needs

Milestone #2 - from May 7th to July 2, 2019

Sprints 6-9

1. Re-evaluate File System definition with Galaxy in mind
2. Specific MGM tools selected and trained (by SMEs)
3. A couple of MGM adaptors completed
    a. ffmpeg
    b. AWS Transcribe
4. Basic workflow mostly working:
    a. With available MGM/adaptors

Milestone #3 - from July 2 to September 10, 2019

Sprints 10-14

1. Data persistence completed
2. Continued workflow work:
    a. Add injection of AMP item to dataset or dataset pair
    b. Validation of inputs and parameters
3. Start defining the normalized data structure
    a. Each MGM has specific fields/types of outputs
    b. Will this be MMIF? - NO, as determined in the July f2f.
4. More MGM adaptors completed with corresponding structured data
    a. Kaldi
5. Human MGMs - defining phase

Milestone #4 - from September 10 to November 5, 2019

Sprints 15-18

1. AMP button to submit a workflow in Galaxy using Galaxy API
2. Design solution for data types in galaxy
3. Retrieving WF results from galaxy
4. User Login UI - initial work
5. More MGM adaptors

    a. INA Speech Segmenter

Milestone #5 - from November 5 to Jan 14, 2020

Sprints 19-23

1. Implement Data types
2. Batch Ingest (without UI)
3. UI Design with UI/UX expert
4. More MGM adaptors
    a. spaCy
    b. AWS Comprehend
5. UI work:
    a. Workflow submission page
    b. Login/signup - finalize
6. Roles and permissions - define architecture
    a. Decision: we will not implement this in the pilot.

Milestone #6 - From Jan 14, 2020 to March 24th, 2020

Sprints 24-28

1. Human MGMs
    a. notification workflow w/ UI
    b. Transcript Editor integration
2. UI work -
    a. Batch Ingest UI
    b. Workflow dashboard - just the design
3. Roles and Permissions
    a. Interaction with Galaxy
4. Conversion of JSON to viewable formats
    a. Transcripts + diarization combined into VTT (still in progress)
    b. NER outputs to CSV
5. More MGM adaptors
    a. Video OCR
        i. Tesseract
        ii. MS Azure
6. Adding dictionaries to NER tools
    a. Spacy
    b. AWS Comprehend

Milestone #7 - From March 24th to June 2nd, 2020

Sprints 29-33

1. UI work -
    a. Workflow dashboard
    b. Applying new UI design to:
        i. Login pages

        ii. Batch Submission page
2. Human MGMs
    a. NER revision tool integration or creation
3. Search features implemented - backend
4. Define strategy for use of HPC at IU and start implementation
5. Implement database backup

Milestone #8 - From June 2nd to August 11th, 2020

Sprints 34-38

1. Finish HPC implementation
2. Modifications to Timeliner as the AMP NER Revision tool
3. More MGM adaptors
    a. Provide a NE thesaurus as input to an NLP tool
    b. Music - not available yet
4. UI work:
    a. Frontend for searches
    b. Workflow Submission pages

Milestone #9 - From August 11th to Oct 20th, 2020

Sprints 39-43

1. More MGM adaptors
    a. Shot Detection
    b. Contact Sheet generation
    c. Program scanning (??)
2. Human MGMs
    a. Segmentation output evaluation tool (Avalon SME)
    b. UI for selection of final outcomes
3. Testing and refinement of MGM results data structure
4. Testing and refinement of User Interface in general
5. Bagging AMP outcomes - frontend and backend
6. Start: Reporting on time for processing each file and each item.

Milestone #10 - From October 20th, 2020 to Jan 26th, 2021

Sprints 44-50

1. Resuming failed Workflows
2. Bug fixes that come up with app usage
    a. by CMs
    b. by Human reviewers
3. MGM Adaptor:
    a. Facial Recognition
    b. Instrumentation detection (??)
    c. Applause detection
    d. Adjust the INA Speech Segmenter galaxy tool

4. Add HPC tools to Galaxy workflows
5. Add support for versioning of MGMs
6. User guides

Milestone #11 - From Jan 26th to April 6th, 2021

Sprints 51-55

1. Support for CMs
2. HPC
    a. Run performance comparisons
    b. Decide whether to use HPC for future passes of the collections content
3. Metadata import into Target Systems
    a. Planning
        i. What information goes where in Avalon?
4. Analytics and Reporting - Planning
    a. Define desired reports
    b. Define desired level of granularity for Analytics
    c. Collect the data
5. MGM Implementation
    a. Applause detection
    b. Forced Alignment
6. Load Testing
    a. Planning and scripts

Milestone #12 - From April 6th to June 30th, 2021

1. Deliverables import into Target Systems (Avalon)
    a. Transcript files (webVTT format)
    b. NER results (CSV format)
    c. Applause Detection (SME format)
2. Generate contact sheets for all content with new parameters
3. Run Music Library content through Applause Detection
    a. And load some of that in Avalon
4. Upgrade Galaxy - and improve packaging of the app as we do this
5. Load Testing:
    a. Run tests
    b. Address performance issues
        i. Queueing/Load balancing strategy may be needed
6. Improve logging and error handling
7. Improvements in configuration settings (config files)
8. Analytics and Reporting
    a. Writing the reports

APPENDIX B. AMPPD GOALS & ANTICIPATED OUTCOMES

Among the original AMPPD project's stated goals were:

"to maximize findability and usability of audiovisual assets by making AMP available to libraries and archives as an open-source software platform with documented APIs that allow flexible integration with institutions' digital content ingest workflows and access systems."

This definition was sufficient to drive the beginnings of the project, with the various project teams focusing on development and metadata generation mechanisms (MGMs). An additional and equally important driver for AMP was Indiana University's need to gather more useful information about their 300,000 hours of digitized audiovisual collections, which in general has very little associated descriptive metadata. However, as the project has progressed, there has been a need to define more precisely what the goals, with regard to the quality of the data output, should be. Questions include:

- What level of precision should the project be aiming to achieve?
- How do you measure "good enough"?
- When have we achieved "good enough"?
- How "structured" can the metadata be? If the answer is "not very," what are the implications of that?

Before attempting to answer these questions, it must be stated that while it would be valuable to answer all of these questions, the reality is that some answers may not be possible, or at least not as specifically and with finality as project participants might wish.

What level of precision should the project be aiming to achieve?

In 2016, research was performed to identify the type of metadata that collections managers at Indiana University were hoping to gather through a project of this nature. The outcome of this research—among other things—was a list of metadata fields organized by "Required," "High Value," and "Others."

Required	High Value	Others
Title	Genre	BPM (beats per minute)*[45]
Name (specifically, the creator)	Type of title	Color information (chroma values)*
Date (date created or date issued)		Color/BW

[45] Fields with * can aid in discoverability, rights management, or MGM automation.

Department (campus unit)		Date (other)
Role (for specific identification of performer, interviewee, director, etc.)		Duration
Rights status		Ethnicity
Primary permissions		Event (e.g., lecture title, basketball game)
Prioritized		Frequency information (audio frequency)*
Geographic (Recording location) (required to determine rights status)		Full text*
Subject		Gender
Identifier (IUCAT numbers, but also shelf numbers, etc.)		Geographic (other)*
Format (CD, open reel, etc.)		Keyword*
Collection (from which the item comes)		Language
Applied permissions		Linked relationship (URL/URI)
Applied permissions note		Music present (binary: present/not)*
Target audience		Music/Speech (binary)*
		Note
		Part/component
		Phonemes (phonetic transcript)*
		Publisher
		Relationship*
		Sound/silent (binary)
		Source (e.g., provenance, donor)
		Type of resource

As this AMPPD project has progressed, the project team has concluded that it is unlikely we will be able to populate all of the values in the “Required” column. As the team understood going into the project, the better the input quality (of the audio and video assets), the better the data output will be. This has been confirmed through application of the MGM technologies thus far in

the project. However, even when the quality of the input is high, MGM output will never be perfect, and it will be extremely difficult to automate decision-making around such things as machines identifying and converting Named Entity Recognition (NER)-produced keywords into LCSH subject terms.

Precision, then, must be measured by the ability of the platform to produce "good enough" structured metadata fields[46], most of which fall into the last column, "Others." At least two fields from the "Required" column will be harvestable in some form, as well. However, in the case of these two fields—"Name (specifically, the creator)" and "Date (date created or date issued)"—the data produced will often be unattributed names or unidentified dates, no matter their role or importance to the audiovisual asset.

It may even be the case that the metadata produced from AMP is not even "structurable"—it could be viewed as "dirty OCR,"[47] containing a blob of identifiable keywords, names, dates, geographic locations, and other values that can aid in further MGM automation, e.g., music vs. speech or silence vs. sound.

Ultimately, the goal of AMPPD is to produce data that is "good enough," which may still require human intervention, either through Human MGMs or work performed after the data is made public. It is data that researchers may be able to identify which (if any), names are creators, dates represent the time of creation, or keywords are subject headings.

How do you measure "good enough"?

The ultimate question is, what is "good enough" data? When does the platform reach the discernable end of improving the metadata? This leads us back to the idea of "dirty OCR." While not the same, a similar way of measuring the quality of auto-transcribed textual content—in this case, transcribed output from audiovisual assets—is the "Word Error Rate (WER)." For contemporary documents produced on computers, and scanned or OCR'd directly from a PDF, "most OCR software provides 98 to 99 percent accuracy."[48] It is likely this will rarely, if ever, be the case for auto-transcribed data produced from transcripts of the audio and video in the historic collections of the New York Public Library and Indiana University. The measurement of WER from audio transcription is telling:

[46] Defined here as data that fits into a specific field based on established rules.
[47] Electronic documents resulting from inaccurate Optical Character Recognition.
[48] Using OCR: How Accurate is Your Data?
https://tdwi.org/articles/2018/03/05/diq-all-how-accurate-is-your-data.aspx

"With an average WER of 16%, Google Speech (Video) is the most accurate ASR engine in our testing. For many audio samples, Google's engine scored a WER well under 10%—as low as 2% for some high-quality audiobook samples."[49] [50]

It is fair to say that "good enough" when dealing with standard, English computer-generated text will, for the near future, be much more accurate than the transcripts output from audiovisual content. Our expectations of "good enough" audio transcription, then, should be tempered.

Having said this, the phrase "good enough" must be defined as precisely as possible. AMPPD team members considered the statement, "AMPPD metadata must be good enough to _____." This exercise prompted the following list, which provides a clear definition of what the AMPPD project is aiming for:

AMPPD metadata must be good enough to . . .

- Help (and not hinder) catalogers describing audiovisual assets
- Increase access to audiovisual assets through search
- Help provide greater context to audiovisual collections
- Increase navigability of audiovisual assets through provision of basic segmentation or structure

When have we achieved "good enough"?

The answer to this question will be highly dependent on the quality of the audiovisual content input into AMP. Each set of audio from the historic collections at NYPL and IU have varying degrees of quality and clarity, as well as very different content ranging from lectures, to interviews, to music performances, to protests and ball games. This will impact the quality of the WER in transcripts directly and meaningfully. This, in turn, will have a likewise significant impact on the output of MGMs applied further in the workflow, including Natural Language Processing (NLP) and Named Entity Recognition (NER). The old adage, "Garbage in, garbage out," unfortunately comes into play here. With that in mind, it is reasonable to say metadata will need to be reviewed on a collection-by-collection[51] basis to ascertain whether AMP has done all it can to produce "good enough" output. The role and value of human MGMs within workflows, performing steps such as transcript correction, will also need to be considered and decided on a collection-by-collection basis.

[49] Which Automatic Transcription Service is the Most Accurate?—2018. https://medium.com/descript/which-automatic-transcription-service-is-the-most-accurate-2018-2e859b23ed19

[50] In many cases during AMPPD testing, Google Speech did not score even as high as 16%.

[51] Assuming that collections of materials generally have the same quality and content.

How "structured" can the metadata be? If the answer is "not very," what are the implications of that?

This leads to the question of structured metadata, which was an inferred goal in the original grant proposal. The project plan is to attempt to create a few structured metadata entities, understanding that these, too, will likely contain some "dirty" data. Ultimately, however, the AMPPD team feels that populating the following entities might be achievable:

- Keyword
- Name
- Geographic Location
- Date
- Full text

There are a number of means for utilizing the data output from AMP. First, it is likely that some institutions will want to remediate this data before adding it to catalogs or digital collections systems. Humans may analyze and create subject terms from the values in "Keyword," "Name," and "Geographic Location." Values in the name field may provide clues to who the creator of the audiovisual asset is. Dates, as well, may provide insight into when an asset was created or edited.

In other cases, institutions may simply grab all of the data about an asset and put it into fields specific to machine-generated data or a combined "Full Text" field. This data may or may not be visible to a researcher,[52] but indexing this data contained within the audiovisual asset in their cataloging or digital collections system could enhance discovery far beyond what metadata is available about the asset.

Ultimately, the data produced from AMP may not be as precise as the AMPPD project team might have hoped at the start of the project. Still, it can be useful—even in its "dirty" form—to enhance cataloging and discovery of the audiovisual assets from which it is produced.

[52] It should be noted that there are also negative effects of indexing dirty data. Inaccurate data could lead to false hits in search, which could be even more frustrating if the researcher isn't able to see why their search term led them to the resource.