AMPPD: Audiovisual Metadata Platform Pilot Development Proposed Activities and Rationale

June 22, 2018 (Public version January 9, 2019)

The Indiana University Libraries, in collaboration with the University of Texas at Austin, New York Public Library, and digital consultant AVP (formerly AudioVisual Preservation Solutions), request $1,251,299 from the Andrew W. Mellon Foundation to support initial development, implementation and pilot testing of an Audiovisual Metadata Platform (AMP) that will enable more efficient generation of metadata to support discovery and use of digitized and born-digital audio and moving image collections. The project will take place over a period of 27 months beginning on October 1, 2018, and will also be supported through substantial in-kind staff contributions from Indiana University. AMP will enable the creation and execution of workflows that link together both automated and human analysis activities, and it will be tested against representative media sample sets from three specific collections, drawn from the collections of Indiana University (IU) and New York Public Library (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.

Project Motivation

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

[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. https://www.weareavp.com/quantifying-the-need-a-survey-of-existing-sound-recordings-in-collections-in-the-united-states/

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

and international digital platforms. The relative paucity of audiovisual content in Europeana and the Digital Public Library of America is testament to the difficulties that the GLAM (Galleries, Libraries, Archives, and Museums) community faces in creating access to their audiovisual collections. As of May 2018, Europeana comprises 53% images and 44% text objects, but only 1.3% sound objects and 2.3% video objects.[3] DPLA is comprised of 52% images and 45% text, with only 0.33% sound objects, and 0.4% video objects.[4]

Another reason, beyond cost, that audiovisual recordings are not widely accessible is the lack of sufficiently granular metadata to support discovery, identification, 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. User personas developed during the pilot grant (see Appendix 1) exemplify the array of discovery needs users have and which are not being met in many cases through the audiovisual metadata produced today.

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 truly derive value from 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.

What is required for full descriptive access to audiovisual objects at scale are a variety of mechanisms (both automated and manual) working together to perform analysis of media and their associated materials (such as transcripts or transcribed information on carriers) in order to generate usable and meaningful metadata that supports discovery, navigation, rights determination, and permissions and access decisions. These mechanisms might include natural language processing, speech-to-text conversion, facial recognition, silence detection, scene detection, music detection, language recognition, manual description, optical character recognition, object recognition, and more. It is not exclusive to automated mechanisms, however. 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, in local and cloud-based systems. The metadata must be compiled, refined, and delivered to a metadata warehouse where it can be harvested by target systems. At the same time, it must remain available to the metadata generation mechanisms (MGMs) for continued and ongoing "cultivation" by the evolving mechanisms'

[3] Europeana. http://www.europeana.eu/portal/en/search?q=
[4] DPLA. https://dp.la/search

technologies and machine learning. In this way, the metadata remains in a constant and active state of refinement.

Two overarching use-cases speak directly to the motivation for AMP:

Collection Managers

Managers of collections of digitized and born-digital audiovisual content have a clear and present need for the ability to generate quantities and types of metadata that are currently out of reach. Generating the metadata they need to effectively manage, preserve, and make their collections discoverable and usable is difficult because there are limited funds with which to support the technological and human resources to generate this metadata, and/or the technical complexity required exceeds the skills within the collection manager's department or institution. AMP intends to address this use case and need by providing a platform that can be deployed and used with minimal cost, an interface for creating complex workflows of metadata generation mechanisms (MGMs) that is simple enough to be used by collection managers, and an overall approach that optimizes human labor and creates more cost effective workflows for generating metadata.

The flowchart in Appendix 2 shows an example of the type of workflow that AMP is designed to be able to create and execute, incorporating automated analysis and extraction MGMs such as scene detection, facial recognition, speech vs. music detection, speaker identification, speech-to-text, and named entity recognition to generate metadata that will then be vetted and improved by human steps to enable end-user discovery and navigation of a previously undescribed video object.

Such a workflow would be created and run by a collection manager following a process such as that shown in Figure 1 below.

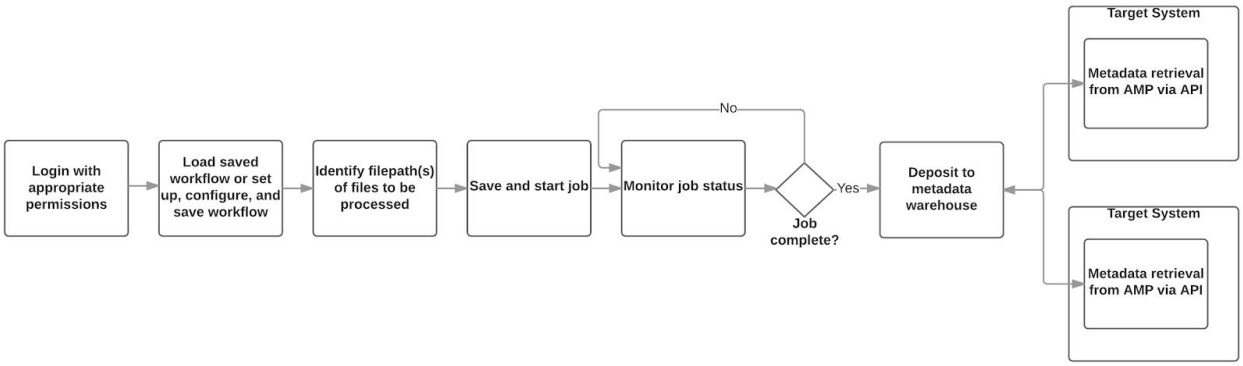

Figure 1. Example collection manager interaction

Users

Users are defined here as any user, including students, faculty, researchers, producers, and anyone else who has an interest or need to find and use audiovisual content held by collection managers for some purpose.

To date, the lack of metadata for audiovisual collections has been a major impediment to the discovery and use of archival audiovisual collections. AMP aims to address this use case by generating significant amounts of structured and time-stamped metadata that can be leveraged for meaningful discovery and navigation of audiovisual content. As an example of this, we can look at a videotape held by Indiana University, containing a recording of an interview with former university president Herman B Wells, which was recently digitized. The current description in IU's media collections portal for this recording consists of the following metadata:

- Description – Radio and Television Services Collection - A Conversation with HB Wells – 40000000784944 – Betacam (31:07)

To demonstrate the value of what is possible with AMP, and the difference to users that are searching for and using content, here is an example of the structured metadata that could be generated with AMP through a combination of leveraging existing data from IU sources and use of MGMs:

- Title – A Conversation with HB Wells
- Title (Type of Title – Alternative) – A Conversation with Herman B. Wells
- Collection – Radio and Television Services
- Department/Unit – Radio and Television Services
- Physical Description – Betacam
- Duration – 31:07
- MDPI Barcode – 40000000784944
- Type of Resource – moving image
- Date Created – 1970/1979
- Date other – 1940/1970
- Music/speech – Speech
- Sound/silent – Sound
- Color/B&W – Color
- Music present – Yes
- Genre – Interview
- Subjects:
    - Indiana University
    - Universities and colleges—United States—Administration
    - Universities and colleges—United States—History—20th century
    - Wells, Herman B.
- Names:
    - Mundt, Bruce (Producer)
    - Petranoff, Robert (Producer)
    - Office of University Relations (Production company)
    - Wells, Herman B. (Interviewee)
    - Clark, Thomas D. (Interviewer)
    - WTIU Bloomington (Production place)
    - Bryan, William Lowe (Referenced)

    - Kinsey, Alfred (Referenced)
    - Roosevelt, Eleanor (Referenced)
    - Rockefeller Foundation (Referenced)
- Geographic (Recording location) – United States
- Geographic other:
    - Bloomington
    - Indiana
- Parts/components – 1 of 1
- Notes – Betacam is transfer from film original based on film leader at start of content. Date range is based on Main Library building being complete and having electricity.
- Rights holder – Indiana University[2]
- Rights status – In Copyright
- Year of Renewal – 2067
- Publication status – Unpublished
- Applied permissions – Campus only
- Default permissions – Campus only
- Language – English
- Keywords:
    - Board of Trustees
    - Faculty
    - University administration
    - University president
    - Music
    - Arts
    - Midwest universities
    - President's spouses
    - 20th-century leaders
- Full Text:
    - OCR from text within video
    - Speech-to-text conversion
- Speaker Identification – [time stamp and name each time a speaker changes, combined with the names field]
- Relation – [Content matching will identify where any portions of content in this video also exist in any portions of other videos within the repository]

The broad results of this amount and structuring of metadata include:

- Discoverability based on individual and combined values for names, subjects, dates, geographic location, and content containing speech.
- Discovery based on full text of the interview, linking to precise location of where a word is said/presented in an interview
- Linking to precise locations for keywords, names, and subjects
- Discovery based on people and topics covered in interview, geographic location, and time period covered in interview
- Identification of underlying music for potential rights clearance needs
- Discovery by color vs. black-and-white

- Discovery and navigation by speaker
- Identification of where specific content is replicated elsewhere in the IU repository
- Ability to communicate and act on permissions associated with rights determination

Using the metadata that exists today, discovery opportunities are extremely limited. A user would likely search for "Wells" or "HB Wells" to find this video along with many others, and then they would need to watch and listen to much of it to find out the contents of the interview.

Leveraging the metadata from AMP, users may conduct searches 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, it was produced at WTIU Bloomington, and the Rights holder is Indiana University.
- 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 not music present.

Beginning to address these two overarching use cases and uncovering the underlying opportunities that exist for organizations and users is a major motivation for this project. An intuitive system that is easy for non-developers and non-technical caretakers of collections to use could change the prospect for future access to hundreds of millions of hours of audiovisual 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.

The eventual goal, then, would be 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.

Previous Work

This proposal is preceded by a workshop and resulting white paper funded by the Andrew W. Mellon Foundation and hosted by IU as part of a planning project for design and development of AMP. The partners leading this planning project were the IU Libraries, University of Texas at Austin (UT) School of Information, and AVP.

The AMP 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 (AMP Project Team Member)
- Kristian Allen, UCLA Library
- Jon Cameron, Indiana University
- Tanya Clement, University of Texas at Austin (AMP Project Team Member)
- Jon Dunn, Indiana University (AMP Project Team Member)
- Maria Esteva, Texas Advanced Computing Center, University of Texas at Austin
- Michael Giarlo, Stanford University
- Juliet Hardesty, Indiana University (AMP Project Team Member)
- Chris Lacinak, AVP (AMP 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 (AMP Project Team Member)
- Felix Saurbier, German National Library of Science and Technology
- Brian Wheeler, Indiana University
- Maria Whitaker, Indiana University

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).[5] AVP gathered information through interviews with collection managers 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 (.xlsx, databases), and/or structures (.xml, .csv, .txt) it resides. Collection managers 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 audiovisual assets in IU's Media Collections Online audiovisual access system based on the open-source Avalon Media System[6] 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, 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.[7]

---

[5] https://mdpi.iu.edu/

[6] Funded in part by grants from the Andrew W. Mellon Foundation and Institute of Museum and Library Services (http://avalonmediasystem.org/)

[7] See Framing Statement Appendix - MGMs and Descriptive Metadata for a diagram of possible MGMs and metadata fields potentially supported by them.

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,
- 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 (in the initial stages) and other systems that offer metadata management and/or discovery related to audiovisual content (in subsequent development cycles). Note that a core concept of AMP is to be usable by any target system. The target systems used in this pilot are serving as a demonstration.

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[8] released in March 2018 that summarizes the output of the workshop and planning project and recommends the next phase of work that is described in this current proposal.

Related Work

Parallel to the work performed at IU, Tanya Clement's High Performance Sound Technologies for Access and Scholarship (HiPSTAS) project[9] at the University of Texas is conducting research on how users can better access and analyze spoken word collections of interest to humanists through:

- an assessment of scholarly requirements for analyzing sound,
- an assessment of technological infrastructures needed to support discovery, and

[8] Dunn, Jon W., 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

[9] Funded by National Endowment for the Humanities and Institute of Museum and Library Services (https://blogs.ischool.utexas.edu/hipstas/)

- preliminary tests that demonstrate the efficacy of using such tools in humanities scholarship.

The HiPSTAS project has produced and documented workflows[10] to show the movement and organization of files in "jobs" for the analysis of large collections of media. The workflows have been tested on collections of cultural heritage audio recordings including field recordings, oral histories, poetry performances, radio programs, and speeches at the UT Austin's School of Information, as well as several other institutions. Output metadata about these files includes genre and speaker identification, among other features.

In addition to the work of HiPSTAS, there have been several open-source and commercial efforts to date that demonstrate the possibilities for computationally assisted metadata generation and improved discovery. This research found that in many cases these systems are fixed, one-way pipelines that are designed to address a single type of content rather than configurable, genre-agnostic systems. For instance, an oral history application may employ two MGMs—a speech-to-text conversion tool and natural language processing tool—in a workflow to process audio files and output a text-based document as a result. While a workflow like this serves as a proof-of-concept for the general approach of AMP, it fails to meet the broader needs defined by user personas and requirements for AMP. These needs call for a system like AMP with the following characteristics:

- Dynamic, flexible, and configurable
- Two-way synchronous communication throughout
- Extensible, scalable, and modular
- Inclusive of human generation of metadata
- Contains conditional logic
- Outputs to a variety of data formats
- Topic or subject area agnostic
- Able to improve accuracy of previously applied MGMs by reinvoking them after additional data has been generated by later MGMs

There are other systems that specialize in a particular type of content or subject, including the multi-institution tool MALACH (Multilingual Access to Large spoken ArCHives),[11] Cornell Lab of Ornithology's Raven,[12] and BBC's Comma.[13] While the level of description these systems generate is extensive and deep for the subjects and content types for which they were built, the focus is narrow relative to the goals of AMP. Also, these systems are not built to handle the scale demanded by AMP or to support the rights and permissions requirements that are fundamental to AMP.[14]

[10] See https://blogs.ischool.utexas.edu/hipstas/hrdr/
[11] https://malach.umiacs.umd.edu/
[12] http://www.birds.cornell.edu/brp/raven/RavenOverview.html
[13] http://www.bbc.co.uk/rd/projects/comma
[14] These comments are not meant to disparage the great work behind these tools or the important work they perform. They are only meant to demonstrate why they are not the right fit for the vision and requirements laid out for AMP.

Most closely approximating the goals of AMP are the commercial platform GrayMeta[15] and the open-source European Union project MiCO.[16] GrayMeta launched at the National Association of Broadcasters convention in April 2016 and was discovered by the AVP members of the AMP project team at the convention one week after finalizing AVP's previous work for IU on description strategy. Upon discovery of GrayMeta and as work continued, members of the AMP project team researched the company, conversed with GrayMeta, and received system demonstrations. The MiCO project was discovered by the AMP project team during interviews with Fraunhofer regarding their automated MGM offerings. In all of our interviews with companies that create MGMs, we provided background about the AMP project and asked if there were any related projects that we should be aware of. As part of this due diligence, Fraunhofer referenced the MiCO project as an EU project that they were taking part in and provided us with recently published resources and a website.

Both of these systems share many attributes of an envisioned AMP system, including the ability to deploy a variety of MGMs, analysis of multiple media types, and storage of resulting metadata in a metadata warehouse. The MiCO project has extensive documentation that is publicly available. However, there is little public documentation available about GrayMeta, which is a licensed product used primarily by two domains (broadcasting and advertising). Thus the findings of our GrayMeta research are based only on system demonstrations and review of the limited marketing materials available, but it was clear from that research that the goals of GrayMeta differ considerably from those of both the MiCO and AMP projects.

A core tenet of AMP is the inclusion of MGMs involving human interaction when cataloging or subject matter expertise is required, for the refinement of automatically generated metadata, and when human feedback supports or drives machine learning. Neither MiCO nor GrayMeta currently incorporates human-interaction MGMs in the integrated way that this is envisioned for AMP.

Additionally, AMP aims to utilize related supplementary documents (e.g., catalog records and transcripts) to augment a media file's metadata. This concept is not represented in either system. Instead, each object in MiCO and GrayMeta is treated as a single information package distinct from all other sources of data.

The target market for AMP is libraries and archives. Based on GrayMeta's public literature and some of the demonstrated features, they are highly focused on broadcast and advertising. The MiCO project is focused on video production environments and animal identification and analysis. This does not mean that their use cannot be extended to other disciplines, but their immediate target markets do influence their current implementation of MGMs and choice of media types and system features. In both cases, the media content on which they focus is contemporary and consistent (broadcast quality), and for that reason is high quality and relatively easy for automated MGMs to work with. Library and archival content, by contrast, is challenging due to the extreme variation across collections in content type, recording quality, recording

[15] https://www.graymeta.com/
[16] https://www.mico-project.eu/

specifications, and subject and discipline areas. The variability that exists in libraries and archives, then, requires specialized system design and performance and is a significant reason for AMP's emphasis on generation and refinement by humans.

Based on planning workshop discussions, there is a strong preference for the use of open-source technologies for AMP. AMP is intended to support an ecosystem that marries commercial, proprietary MGMs with open-source MGMs (although open-source components are strongly preferred where possible). For this reason, use of a commercial, proprietary, "black box" system to fulfill many, or all, of AMP's functions was deemed undesirable by planning workshop participants.

Use of proprietary systems becomes particularly problematic when considering the paradigm shift that AMP, and platforms like AMP, require in thinking about ownership. Traditionally, the value derived from the work to describe content has been the text output itself. Whether the description was performed in-house, through a service, or was performed by humans or automated processes, value and ownership remained on the data that was produced through the descriptive process. The deliverables have always been static text, in .txt, .doc, .docx, or .pdf documents, or structured in .xml, .csv, .json, MARC or other formats. AMP shifts this paradigm by storing the metadata outputs in a metadata warehouse where it will continue to be cultivated, groomed, and refined as MGMs evolve or as additional MGMs are integrated into AMP.

Machine learning will be leveraged over time as new algorithms are incorporated into AMP and humans continue to refine and provide feedback to improve the accuracy and performance of the MGMs. Amazon Web Services summarizes machine learning (ML) training as "...providing an ML algorithm (that is, the learning algorithm) with training data to learn from. The term ML model refers to the model artifact that is created by the training process."[17] The evolving machine learning algorithms and models and the human resources that help to create a smarter machine will become one of AMP's most valuable assets. With this in mind, ownership of the "machine" and the transparency of its functions becomes as important as the outputs themselves. It is critical that a library or archive own and benefit from the algorithms, models, and human resources that continuously build a smarter machine. AMP functionality must be transparent to support institutions' roles as stewards of collections they are charged with preserving, and to ensure the data they are producing is authentic and trustworthy to the greatest extent possible. Such transparency would not necessarily benefit a commercial entity, as it would expose the core of the system's value. Institutions that value—and indeed trade—on openness, trust, and being as unbiased as possible, would not have access to the inner workings of the commercial system, making it difficult to fulfill their missions. This would effectively minimize the returns on their own human and financial investments and metadata quality, and the trust in its authenticity over time would suffer.

This proposed AMP pilot project will serve to validate that an open-source workflow system can be built to apply automated and manual metadata generation steps to produce metadata for audiovisual materials that previously did not exist and that this metadata can be successfully

[17] Amazon Web Services. "Training ML Models."
http://docs.aws.amazon.com/machine-learning/latest/dg/training-ml-models.html

delivered to and used within target discovery and access systems. This work will inform future plans to fully build out the AMP system for use and generate cost metrics for the AMP-based approach that can be compared to fully manual processes.

Work to be Done

This project will have three primary facets: implementation, development, and integration; collection partner collaboration and performance of the pilot; and project meetings and reporting.

Implementation, Development, and Integration

Implementation

The platform architecture workshop conducted during AMP's planning phase resulted in a conceptual model and technical architecture for AMP. This is detailed in the *Technical Approach* section below. For each component in the architecture that has suitable products available, candidates were identified by participants in the workshop held as part of the pilot grant. The process of implementation will require a more critical vetting of component candidates to gain a deeper understanding of exactly how these candidates fulfill the role of their associated components. This will involve research, setting up test instances of candidates, compatibility and performance testing, troubleshooting and configuration, comparative analysis, and final selection. The development team will lead the implementation effort, reporting back to the main project team with salient findings in order to make final decisions on the architecture and candidate systems.

Throughout the planning phase, the AMP project team studied the MiCO project (discussed earlier under *Related Work*) and met with the MiCO project team on several occasions. Funding for the MiCO project ended in 2017, and while the project was a multi-organizational effort, the Fraunhofer Institute for Digital Media Technology (Fraunhofer IDMT) played a lead role. Because of Fraunhofer's role in that project and their interest in AMP, our continued discussions regarding prospective integration of MiCO project code were primarily with Fraunhofer. By the end of our technical planning phase, we had determined that it would be technically and financially advantageous to utilize MiCO project code to the greatest extent possible within AMP. With MiCO code reuse in mind, we have integrated a team from Fraunhofer into this proposal. We expect that most of their work will be in the first 15 months of the project, serving in an advisory capacity to ensure that we implement and utilize MiCO project in the most efficient and effective way possible. Specifically, in the Implementation, Development, and Integration phase, Fraunhofer will participate as follows:

1. Remote participation in technical planning and working sessions consisting of multiple deep dive meetings into particular aspects of the MiCO project code deliverables.
2. Remote participation at monthly meetings in which the core AMP project team will report and seek feedback on matters relating to the MiCO project code deliverables.
3. Informal ad-hoc communication (e.g., Slack channel, screen share, video conference meetings) with the AMP developer team to respond to specific questions regarding lower-level coding and implementation matters.

In this way, AMP team will have the necessary support to fully leverage the MiCO project code deliverables.

Development

The *Technical Approach* section mentions the anticipated need to develop components where there are no suitable candidates available. The AMP project team currently believes that the storage management system and some components of the workflow management system will need to be developed. The implementation work will offer additional vetting and due diligence to confirm or revise current thinking and refine our understanding of requirements for systems that need to be developed.

The understanding gained from the implementation effort will be combined with documented requirements and user personas produced as part of the planning phase. These will be further supplemented with detailed requirements regarding deliverables for Collection Partners that will be gathered as part of conversations taking place early on in the project. These discussions will also cover the content they are submitting, the fields they would like in order to support their users and use cases, and the target systems they would like to populate as the result of this pilot project, all of which will yield insights into the MGMs that we will need to utilize. These MGMs will consist of multiple types, including manual MGMs (work performed by humans) which may require the development of user interfaces for human input. Manual MGMs might include visual metadata clean-up before further processing, review or proofreading of metadata in-between application of automated MGMs, population of metadata fields for which automated MGMs are not sufficient, and evaluation of metadata quality and completeness by a human to decide if further automated MGM processing should occur. The University of Texas at Austin will engage in testing of existing candidate automated MGM implementations that cover various areas of needed metadata generation functionality, to help identify specific MGM implementations for integration with AMP.

Collectively, these functional and technical requirements, use cases, and MGM information will serve as the basis for beginning the development effort. For this project we will employ the Agile Scrum project management methodology, with the project team working in a tightly integrated and highly collaborative way. AVP will be taking on the roles of Project Manager/Scrum Master, Developer, and Subject Matter Experts. IU will be taking on the roles of Product Owner, Lead Developer, Developer, and Subject Matter Experts. Combined, these roles will make up the core project team with Fraunhofer playing the role of contributors and advisors at pertinent points in the process. For more detail on roles and responsibilities, see the *Organizational Structure* section below.

The development work will be organized into two-week sprints. At the end of each sprint, a demonstration of the work performed over the previous two weeks will be given to the Product Owner and interested stakeholders. An important aspect of the Scrum methodology is that demonstrations consist of showing actual working software representing requirements. The requirements demonstrated are then either accepted or rejected by the Product Owner. Following the demonstration, another subset of requirements is prioritized for the two weeks of work ahead, and those requirements are discussed to flesh out the details. The Developers work together to

figure out the fine grained details of which developer will take on which tasks, although they will work in a collaborative way throughout. Developers will brainstorm solutions together, ask each other for feedback on work performed, and perform quality assurance for each other. Finally, the sprint meeting ends with a retrospective in which the core team discusses what is, and is not, working well in an effort to continuously improve. The Scrum Master and Developers meet on a daily basis for a brief meeting, referred to as the daily standup meeting, in which the agenda consists of reporting what was completed the previous day, the work plan for the day ahead and a discussion of any obstacles that need to be addressed.

The following systems will be used to support the Agile Scrum development process and the overall work of the project:

- *Project Management:* Jira will serve as the system of record for documentation about work to be performed, current work status and past work performed. This will be central to the effort, used for creating, managing and communicating about requirements, stories and timeline.
- *Communication:* Slack and email lists will be used for informal written communication amongst team members. Video conferencing tools such as Zoom or Google Hangouts will be used for meetings amongst the project team, with collection partners, and with other stakeholders.
- *Source Code Repository:* GitHub will be used as a repository for storing and managing the source code and making it available to other interested parties.
- *Documentation:* GitHub will be used for technical documentation, consisting of both inline documentation within code as well as narrative documentation regarding prerequisites, installation and configuration, and other "back-end" documentation. In addition, GitHub will be used for "front-end" user documentation, consisting of information on how to use different aspects of the system.

Following these protocols and utilizing these systems, new components will be developed for incorporation into the architecture as needed.

Integration

The development process described above will come to include integration work at the point at which the development effort matures. This phase will focus on:

1. Integrating the newly-developed components with the off-the-shelf components such as the database, messaging system, and job queueing systems, to complete the platform architecture and ensure that it operates as a cohesive, unified platform.
2. Integrating selected MGMs with the platform. Selected and developed MGMs will need to be able to communicate with the platform and exchange data. MGMs will consist of manual, automated, and hybrid MGMs, as well as locally run and cloud-based MGMs, and free and paid MGMs. Each of these types will require its own integration mechanism, protocol, and effort.

Integration work will blend into the development effort and utilize the same Agile Scrum project management methodology described above.

Collection Partner Collaboration and Performance of the Pilot

Collection Partners will be engaged toward the beginning of the project to ascertain requirements, use cases, details about the content and metadata they are submitting, and requirements for deliverables from the pilot project based on their target system ingest/import specifications.

Two Collection Partners will be identified from within IU, and NYPL will participate as an External Collection Partner.

Potential IU Collection Partners and collection use cases include:

- Cook Music Library collections: Recognized as one of the largest academic music libraries in the world, the William and Gayle Cook Music Library's audio and video collection contains more than 165,000 recordings. The collection is particularly strong in the area of opera, and features among its special collections the Jussi Bjoerling Collection (the world's largest collection of recordings by the Swedish tenor, about 3,000 items), and the Ross Allen and Alvin Ehret collections of vocal recordings (37,000 recordings of operatic and vocal repertoire, including virtually all complete operas recorded between 1950 and 1975, many of them unique or rare in the United States), original audio and video recordings of IU Jacobs School of Music performances and special collections of commercial music audio recordings

- Archives of Traditional Music: With over 100,000 recordings, including more than 2,700 field collections, the Archives of Traditional Music is one of the largest university-based ethnographic sound archives in the United States. Its holdings cover a wide range of cultural and geographical areas, vocal and instrumental music, linguistic materials, folktales, interviews, and oral history, in the form of nearly 7000 wax cylinders, 4600 lacquer discs, 2625 aluminum discs, 250 wires, 18,000 open reel tapes, 7500 audio cassettes, 911 films, and 1500 video recordings. Documentation for field collections potentially useful as supplemental inputs to AMP include researchers' field notes along with "index sheets" in the form of typewritten or word processor documents detailing recording contents

- University Archives, including video recordings of significant university events, guest lectures, and interviews with faculty, students, and administrators

- Center for Documentary Research and Practice, including the IU Oral History Archive sound recordings and transcripts created as part of the archive's mission to preserve, collect, and interpret 20th-century history through the medium of first-person testimony.

Potential NYPL collection use cases include:

- Joffrey Ballet Collection: Now located in Chicago under Artistic Director Ashley Wheater, The Joffrey Ballet is known for technical brilliance and a repertoire that mixes classic ballet with newly commissioned work. Founded by Robert Joffrey in 1956 in New York, The Company has long expressed a diverse and inclusive perspective on dance and

was known to push creative and technical boundaries. The NYPL Archives and Special Formats Processing units have identified the Joffrey Ballet performance footage as a prime collection candidate for the AMP project.

- AIDS Activism Videotape Collection: The AIDS Activism Protest collection consists of original videotapes and masters of completed works documenting the grassroots response of artists and activists to the AIDS epidemic between 1985 to 2000. Much of the collection focuses on the protest actions of ACT UP and other activist organizations, including demonstrations at the White House, Centers for Disease Control, the Food and Drug Administration, National Institutes of Health, City Hall and St. Patrick's Cathedral in New York City, as well as protest video from Testing the Limits records (a video collective formed in New York in 1987 to document AIDS activism) and the Gay Men's Health Crisis records (America's oldest AIDS organization, formed in 1982, serves to educate the public about HIV/AIDS, provide care services for People with AIDS (PWAs), and advocate at all levels of government for fair AIDS policies).

- Jeff Kisseloff audio and video oral history interviews: Jeff Kisseloff is an American writer and oral historian. His first book, *You Must Remember This: An oral history of Manhattan from the 1890s to World War II* (1989) grew out of a supplement he wrote for the *Chelsea Clinton News* newspaper where he interviewed longtime Chelsea and Hell's Kitchen residents. He expanded his interviews to include over 150 interviewees from across Manhattan, framing his interviews by neighborhood. His next book, *The Box: An oral history of television, 1920-1961*, was published in 1995. A longtime baseball fan, Kisseloff began conducting interviews in 1995 for a book on the integration of American baseball. He interviewed baseball players, managers, and journalists but the project was never completed. Kisseloff wrote two young adult books on baseball, *Who is Baseball's Greatest Hitter?* and *Who is Baseball's Greatest Pitcher?* From 1999 to 2000, Kisseloff interviewed social and political activists for his third book, *Generation on Fire: Voices of protest from the 1960s*, which was published in 2007.

The potential IU Collection Partners and collection use cases were selected because they represent large collections to which application of mass approaches to metadata creation could result in significant time and cost savings over traditional fully-manual transcription and cataloging approaches, and because they have content readily available in digital form that was created through IU's Media Digitization and Preservation Initiative (MDPI).

NYPL was selected based on their interest and commitment to participate, along with the ability to contribute collection materials that are complementary to collections available from within IU. Like IU, NYPL has embarked upon a mass digitization effort to convert its physical collections to digital form and thus has similar needs around scaling up description and rights work in order to provide digital access. This diverse data set will allow us to fully test and assess AMP against a range of collection types and potential users.

Collection Partners will be asked to submit up to 100 hours of content each along with any relevant existing metadata, possibly including spreadsheets, library catalog records, related

written documents/programs, transcripts, indexes, shot logs, and more. It is also feasible that supplemental materials to support machine learning may be requested/submitted as well. For instance, photos containing faces along with names of the people in those photos may be used to support facial recognition, when such a resource exists. Once the media and metadata are identified for each Collection Partner, the partners will have approximately nine to twelve months to prepare that content for delivery.

Before media and metadata are received, the AMP project team will design the workflows and pilot project test plans for each Collection Partner. The test plan will define the key questions to be answered and primary benchmarking metrics that will be sought through the pilot project. In addition to this, the test plan will identify the test methods to be used and how the outcome of those tests will be measured. For instance, one key question we would like to answer with this pilot is the relationship between resolution and performance across a variety of MGMs in order to ascertain data that will help with optimal bandwidth utilization. In other words, it is apparent that bandwidth will be one of the main challenges. Because of this, it is ideal to send as little data as necessary. Performing a test in which varying resolutions (i.e. high-resolution preservation master, medium-resolution mezzanine file, low-resolution access copy) are sent to a variety of MGMs to analyze the variance in quality of the resulting outputs provides meaningful data in this regard.

Workflow designs and test plans will be documented for each Collection Partner, resulting in multiple workflows per partner needing to be configured in the AMP workflow management system. This will include the selection of MGMs, configuration of MGMs, ordering of MGM inputs and outputs, setting up billing parameters for paid MGMs, identifying the storage location(s) containing the media and metadata, and performing a series of tests to assure that the systems are working properly.

The cost and human resources required for use of MGMs in the processing of content for each Collection Partner will be calculated before processing is performed to ensure that there are adequate resources available to perform multiple rounds of processing. Workflows and MGM usage may be configured to most effectively utilize available funds and resources if necessary. While collections will be selected to represent a diversity of use cases, the workflows are modular, so all or a portion of one workflow could be reused for another Collection Partner's workflow in the future.

After the workflows and test plans are completed and used for processing each Collection Partner's test collection, the data will be analyzed to determine what metadata was generated, what value was added from that generation activity, and what changes to MGMs or workflows would be useful to include in the next test. In addition to this evaluation, the results will be reviewed against the test plans described above speaking to benchmarking performance and quality metrics, cost metrics, and other key criteria identified as part of the test plans. From this work, preliminary plans will be drafted regarding the next round of processing and testing. The results will also be shared with the Collection Partner, and a meeting will be set up to review and discuss the results in order to receive feedback and input on the output and the next round of

processing and testing. At least three rounds of processing, testing, and review will be performed with each of the Collection Partners.

To the extent that it is feasible, each testing round will conclude with populating target systems, which in the case of the IU Collection Partners will be Avalon Media System. For NYPL, the target system will be a combination of ArchivesSpace as an archival descriptive management platform and the NYPL Archives Portal for public access. The core project team, including metadata subject matter experts, will work with the Collection Partners to evaluate the effectiveness and usability of generated metadata within the target systems.

Project Meetings and Reporting

In-person project meetings will take place throughout the project. The primary events taking place under this umbrella are the kickoff meeting and two other in-person project meetings, all to be hosted by IU in Bloomington, Indiana. Draft agendas for in-person meetings may be found in Appendix 3.

- The kickoff meeting will be used to ensure that all project participants are in alignment with regard to project scope, process, responsibilities, roles, timeline, and deliverables (including AMP, itself, and the processed metadata migrated into target systems). This meeting will also serve as an opportunity to gather momentum and support team building through face-to-face interaction. This is the only meeting in the project in which all project participants from IU, AVP, UT Austin, Fraunhofer, and NYPL will attend in person.

- The second project meeting will take place at the end of year one, which will be a critical transition point from the development and implementation of AMP to using AMP in the processing of media and metadata from Collection Partners (i.e. production). This meeting will include IU, NYPL, AVP, and UT Austin participants and will serve as an opportunity to regroup as a team and aid in the transition from one phase of the pilot project to the next.

  Based on team members' past experiences with projects that include transitions from development to production, it can be challenging to create clear lines of delineation between phases, leading to communication issues among project stakeholders, loss of productivity, and lack of clarity. In this meeting it will be important to discuss winding down the development and implementation phase of the project and ramping up the production phase of the project with all stakeholders at the table to help ensure common understanding and cooperation. For winding down the development and implementation phase it will be necessary to identify and clarify outstanding development and implementation issues, implications of these outstanding issues and any associated obstacles to the production phase, paths and timeline to resolution, contingency plans, communication milestones and mechanisms, and assignment of roles and responsibilities. For ramping up the production phase it will be necessary to review the overarching plan and approach, discuss milestones and timeline, review salient details of content quantities and qualities for each content partner, confirm roles and responsibilities, confirm logistics

of exchanging media and metadata, identify any potential issues with completion of the plan, and discuss possible solutions and contingency plans. Working as a cohesive project team with all stakeholders at the table and engaged on creating an effective plan for transition should greatly mitigate the risk of typical transition pitfalls.

- The third and final project meeting will take place at the end of the project. This meeting will be used as a demonstration of the system and pilot project results to all project participants including representatives from IU, NYPL, AVP, and UT Austin. It will also be used as an opportunity to collect feedback on the project and to discuss, prioritize, and document potential next steps.

Following the final project meeting, the AMP project team will prepare a white paper detailing what took place in the pilot project, results and findings, effectiveness and usability of generated metadata, descriptions of the deliverables and how to use AMP, and where the deliverables and associated documentation can be found.

Resources Required

The IU Libraries plan to devote significant existing staff resources to project direction and product management and also to contribute staff resources in the areas of collections expertise, metadata analysis, IT systems engineering, and system administration. However, additional resources are needed from the Foundation to support a number of needed functions, including project management, software development, manual data capture, and both local- and cloud-based infrastructure for running the AMP system and MGMs. Specifically, the project requests funding in the following areas, which are further detailed in the Organization Structure and Budget Narrative sections of the proposal:

- Staffing, consulting services, and subcontracts:
    - Salary and benefits for two software developers to design, code, and test the AMP system
    - Consulting services from AVP to support project management, AV and MGM subject matter expertise, and software development
    - Advisory support for the MiCO Platform from the Fraunhofer Institute for Digital Media Technology in Germany
    - Audio MGM subject matter expertise and consulting from the University of Texas at Austin School of Information
    - Labor costs for the participation of New York Public Library as External Collection Partner
- Equipment and supplies:
    - Virtual machine costs at IU for testing and running AMP and local MGMs
    - Costs for use of external cloud-based MGM service providers
- Travel and meetings:
    - Travel and hosting costs to support project meetings and promotion of AMP project work at relevant conferences

Availability of Results for the Broader Community

The outcome of this project will be an open-source application, released under an Apache 2.0 or similar non-viral open-source license and made available via GitHub. The application may be locally hosted by any institution with the infrastructure to support it. Additionally, AMP project staff will report on project results in a white paper, made available through IU's institutional repository IUScholarWorks, and through presentations at appropriate library and archives conferences, potentially including the Digital Library Federation Forum, Coalition for Networked Information, Association of Moving Image Archivists, Association of Recorded Sound Collections, and others.

Project Organization

The project will be led by the Indiana University (IU) Libraries, under the direction of Jon Dunn as principal investigator and project director, with significant support from AVP, which will serve as a consultant to IU on the project. The software development, release, and testing process will be carried out by a development team using the Agile Scrum software development project management methodology, an approach that has proven successful for IU and AVP on multiple prior projects, including Mellon-funded software projects such as the development of Avalon Media System (IU and Northwestern University) and version 2.0 of the Archival Management System for the American Archive of Public Broadcasting (WGBH Boston and AVP). The team will make use of daily standup meetings and biweekly development review and planning meetings, as well as semiannual in-person meetings to review status and conduct near-term and strategic planning.

AVP will work on the project in three major areas: 1) logistical management of the project, including serving as Scrum Master for the software development process; 2) contributing to AMP software architectural design and development; 3) providing expert advice and analysis on workflow design, metadata implementation, and choice/implementation of MGMs.

An advisory board of six to eight members will be established for the 27-month duration of the project to provide advice in the areas of system architecture and design; choice and application of machine learning-based MGMs for audio, video, and natural language processing; user needs; metadata; and other aspects of AMP system development. The advisory board will meet by phone or videoconference at least three times over the course of the project. Members will be invited from amongst the participants in the planning phase of AMP, possibly supplemented with invited experts from additional areas, including natural language processing and video content analysis.

Existing Staff

Jon Dunn, Assistant Dean for Library Technologies, will serve as Principal Investigator and Project Director (.20 FTE), responsible for overall programmatic direction and financial management of the project, as well as overseeing the involvement of the project's advisory group. He has over twenty years of experience in the development of digital library software technologies and has served as principal investigator or project director on numerous grant

projects funded by IMLS, NEH, and the Andrew W. Mellon Foundation, including the IMLS-funded Variations3 project, IMLS and Mellon-funded Avalon Media System project, and Mellon-funded Integrating Licensed Library Resources with Sakai project.

Maria Whitaker, Head of Digital Media Software Development, will serve as Scrum Product Owner for the project (.30 FTE), responsible for managing the overall set of system requirements and translating them into user stories for work by the development team, and will also manage the IU software developers. She is a Certified Scrum Product Owner, with over thirty years of experience in systems analysis and software development. She has worked in the IU Libraries as software development manager and Scrum Master for Avalon Media System since 2015. Prior to that, she worked in IU's central IT organization in multiple roles, including serving Product Owner for the university's Human Resources Management System.

Julie Hardesty, Metadata Analyst (.20 FTE), will perform metadata analysis and design work for the project, including assisting with design of data structures for the metadata warehouse and target system export functionality. She has extensive experience working with metadata for digital collections held and managed by IU Libraries, establishing standards to use and requirements for discoverability, access, and sharing. She works with metadata for audiovisual materials on projects including Avalon Media System, HydraDAM2, and the IU Media Digitization and Preservation Initiative and has been active on metadata activities within the Samvera Community.

Brian Wheeler, Senior Systems Engineer (.15 FTE), will serve as a technical consultant on systems architecture and implementation. He has been the lead architect and developer of the hardware and software systems for automated quality control, transcoding, and post-processing of audio, video, and film digitization outputs for IU's Media Digitization and Preservation Initiative, which on average successfully process and move over 30 terabytes per day of content.

Jon Cameron, Digital Media Service Manager (.20 FTE), will be the primary interface between the AMP project team and Collection Partners regarding selection of appropriate test collections and facilitation of transfer of collection files to AMP. He currently serves as co-product owner for Avalon Media System and service manager for the production Media Collections Online instance of Avalon at IU, and he is also involved in designing and facilitating workflows for providing public access to audio, video, and film items digitized through the IU Media Digitization and Preservation Initiative.

Thomas Whittaker, Head of Media Cataloging (.20 FTE), will contribute to the design of workflows for the selected test collections and supervise the student hourly staff hired to work on manual metadata generation steps. He has over ten years of cataloging experience with expertise in film and audiovisual formats. As Head of Media Cataloging, he is actively engaged in setting cataloging policy and developing the workflows necessary to support the discovery and access of the IU Libraries' film and audiovisual collections.

Grant-Funded Staff

A **Senior Programmer/Analyst** (1.0 FTE) and **Programmer/Analyst** (1.0 FTE) will be hired by IU using grant funding, to carry out feature design, coding, and testing for AMP.

Student hourly staff will be hired by IU using grant funding to assist collection managers in selection of materials and retrieval of files for use within AMP and to provide other assistance as needed to the project. Student hourly staff will also be hired to complete metadata checking/creation steps as part of manual MGMs.

Consultants

AVP is a consulting and software development firm with offices in New York, Wisconsin, Wyoming, and Florida. Founded in 2006, AVP provides services to clients globally to help overcome the challenges faced in the preservation and use of data. With a strong focus on professional standards and best practices, open communication, and the innovative use and development of technological resources, AVP uses its broad knowledge base and extensive experience to help clients from a variety of sectors efficiently and effectively ensure that content is manageable and accessible for the long-term.

AVP has a successful history of creating open-source and freely available software for the library and archives community.

AVP was the original developer of the Archival Management System (AMS) for the American Archive of Public Broadcasting from 2012 to 2014 and is currently working on the Mellon-funded development of the AMS 2.0. AVP's work for the American Archive in 2012 led to a related project with the Flemish Institute for Archiving (VIAA) to develop another version of the AMS in order to manage a similar effort involving the digitization of hundreds of thousands of hours of content across more than one-hundred organizations.

In addition to the AMS, AVP has developed many free, open-source applications for the community, with support from partners and independently. These applications can be found at https://www.weareavp.com/products/.

AVP, with funding from IU and the Mellon Foundation, has also been involved in AMP since its inception and has been responsible for much of the underlying analysis and concepts, most of which predated the existence of any public information regarding MiCO and GrayMeta. AVP's experience and track record with software development, serving the library and archives community, and AMP demonstrate a unique suitability as a partner on this endeavor.

Consultants from AVP will serve in a number of roles within the project team:

Amy Rudersdorf, a Senior Consultant with AVP, will act as project manager, Scrum Master, and metadata subject matter expert. She trained in Agile Scrum project management and has been involved in AMP since its inception. She has taught graduate courses in metadata, regularly advises on metadata modeling for clients, and prior to her work at AVP developed data strategies and processes and coordinated a national network of institutional partners for the Digital Public

Library of America (DPLA). She has worked extensively with Dublin Core, MODS, EAD, EDM, PREMIS, and MARC, and worked in tandem with DPLA institutional partners to prepare their metadata to transition to the linked-data ready DPLA metadata application profile.

Adeel Ahmad, Senior Software Developer and Architect, will serve in the role of developer. He has experience leading a broad range of projects encompassing web-based, mobile, and desktop applications. Although his experience with varying technologies is wide-ranging, Adeel's particular area of expertise as a developer is in web applications, server configuration and administration, and API development. Adeel is also AWS-certified, bringing knowledge and practical experience utilizing a broad range of cloud-based services and technologies that are relevant to this project. Adeel was involved in the technical planning phase of AMP and has served as the Lead Developer on projects including the American Archive AMS (1.0 and 2.0), VIAA AMS, MediaSCORE/MediaRIVERS, AVCC, Exactly, Fixity, Catalyst, and MDQC. These efforts have involved extensive and in-depth work focused on the management, migration, and transformation of metadata and media for many millions of objects. This has required the ability to work creatively at scale, with thorough documentation, and significant quality assurance and control protocols in place. Adeel Ahmad's work on the Mellon-funded American Archive of Public Broadcasting project at WGBH is scheduled to end by the end of 2018. Adeel's work on AMP, as currently planned, will begin in January 2019.

Chris Lacinak, President and Senior Consultant, will serve in the role of subject matter expert in regard to audio, video, and metadata formats, tools, and workflows. Chris has served as a subject matter expert in the analysis, development, and reporting behind AMP since its inception. Chris has extensive experience in audiovisual signal processing, metadata generation and management, digital asset management, and software development, implementation, and integration. Chris has been involved in the development of OHMS (Oral History Metadata Synchronizer) and other applications. Chris served as an Adjunct Professor in the New York University Moving Image Archiving and Preservation Masters program from 2005 through 2012, designing the curriculum for and teaching five courses. Chris is actively involved in the creation of standards and best practices through standards bodies and other professional organizations. He also speaks, teaches and advocates on behalf of libraries and archives globally. Although originally scheduled to serve as the Scrum Master on the American Archive of Public Broadcasting project, Chris Lacinak has been replaced by Kara Van Malssen in this role and Chris is serving in a limited role as a subject matter expert, leaving ample time to commit to the AMP project.

Staff from **Fraunhofer Institute for Digital Media Technology** (Fraunhofer IDMT), an applied research institute working in the area of audiovisual media, will provide consulting and assistance on the project team's adaptation of the open-source MiCO Platform for use within AMP as the basis for its workflow engine and service orchestration capability. As original developers of the MiCO Platform, Fraunhofer IDMT is uniquely situated to assist in this project.

Subcontractors

Tanya Clement, Associate Professor in the Department of English at the University of Texas at Austin, along with a part-time student assistant, will assist with evaluation and testing of candidate metadata generation mechanisms. Her primary area of research centers on scholarly

information infrastructure as it impacts academic research, research libraries, and the creation of research tools and resources in the digital humanities. She is the PI for the HiPSTAS project. The **University of Texas at Austin** will serve as a subcontractor on the project.

New York Public Library will also be engaged as a subcontractor under the direction of **Melanie A. Yolles**, Head, Archives Unit, with grant funds used to offset labor costs for selection and contribution of digital audio and video files to the project as well as participation in development and refinement of workflows and test plans and evaluation of results.

Technical Approach

The technical approach chosen for AMP development has been informed largely by the output of the platform architecture workshop conducted during the project's planning phase. Over the course of the workshop, a high-level architecture was collaboratively defined for AMP by workshop participants, with the goal of meeting desired requirements for functionality, configurability, ease of use, and flexibility in adapting to new workflows and MGM implementations. This architecture is shown in figure 2.

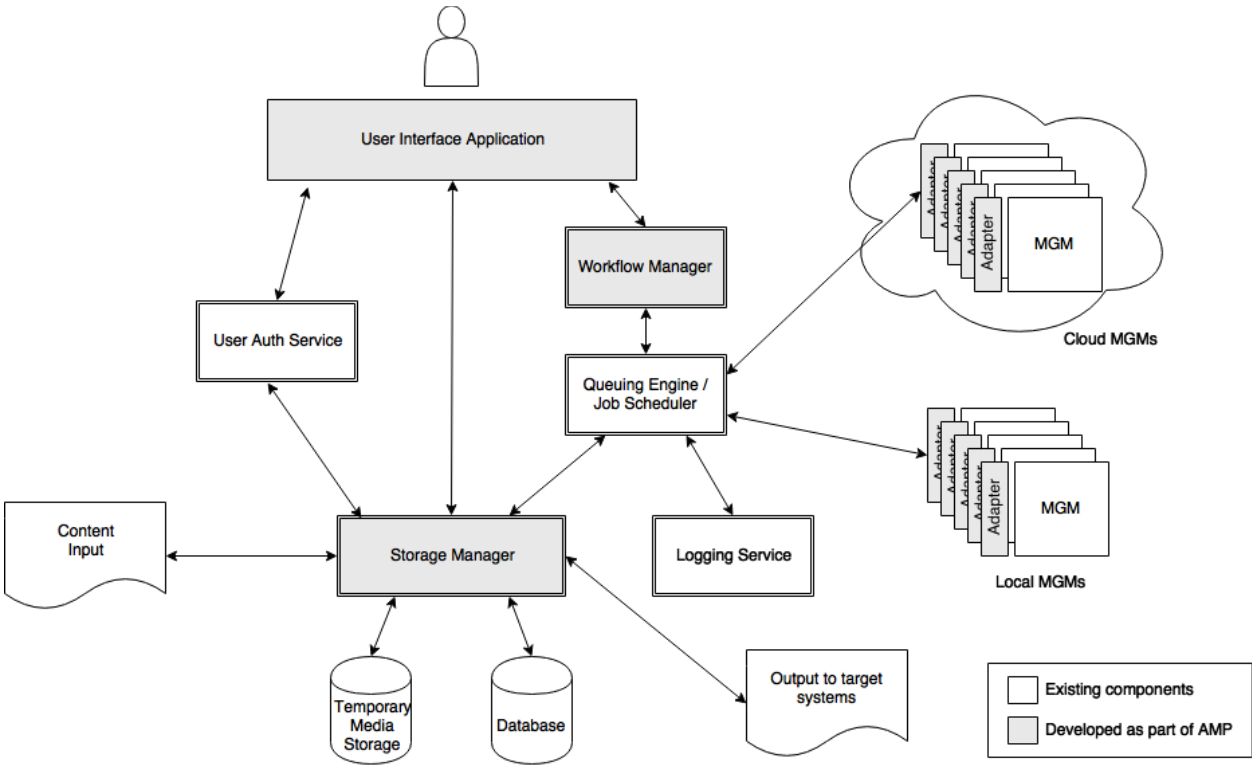

Figure 2. Proposed high-level AMP system architecture

The architecture addresses the stated high-level requirements in the following ways:

- Functionality: The architecture contains the system components necessary for collection managers 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.

- Ease of use: The User Interface Application (UIA) component speaks most directly to the ease of use. The UIA is intended to allow 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. As MGM technologies evolve, the AMP architecture will be able to incorporate these changes, allowing users to leverage new technology capabilities in their workflows.

The AMP architecture will involve a combination of existing software components and new components to be developed. Preliminary options for existing components have been identified, but further work will be required on the part of the development team to make final selections.

The project plans to leverage the architecture and code of the MiCO Platform, developed as part of the EU-funded Metadata in Context (MiCO) project, as much as possible in developing AMP, particularly including the MiCO Broker, which supports orchestration and invocation of metadata extractors (similar to the AMP concept of MGMs) as components of complex workflows. Members of the MiCO project team at the Fraunhofer Institute for Digital Media Technology will be engaged during the project as consultants to assist in implementation of MiCO and its integration within AMP.

Programming language: Given that the bulk of the MiCO Platform has been developed in Java, the AMP project plans to use Java for development of its core functionality. User interfaces will be developed using a client-side JavaScript framework such as AngularJS in conjunction with HTML5 and CSS.

Database: A database is required to serve as a data warehouse for storage of both intermediate and final metadata outputs of MGMs and MGM workflows. Given the variability of input-output format structures, a database structure that can flexibly accommodate both key-value pair and hierarchical data is needed, making a NoSQL database likely preferable to a traditional relational database. Potential NoSQL options identified include Mongo DB, Cassandra, Couch DB, Redis, and Level DB.

Storage Infrastructure: Because of the large size of audio and video media files, a design goal of the system is to minimize network movement and copying of AV data as much as possible. The system will support submission of content via URIs referencing Web-accessible content or content available on a filesystem shared with the machine on which the AMP platform components (and possibly MGMs) are running. Individual MGM steps may generate intermediate transformations of metadata that require temporary storage.

Storage Manager: Storage and retrieval functions for content and metadata will be mediated through a storage management API to enable adaptation to various mechanisms for content input and temporary storage (e.g., filesystem, S3, HTTP) and to different database systems in the future.

Job Queuing: A key component of the system architecture is an engine that can queue, execute, and track tasks performed by MGMs when executing workflows across items in a collection. Potential queuing and messaging systems identified include RabbitMQ, ActiveMQ, MQTT, Apache Kafka, AWS SNS, ZeroMQ, and Apache Thrift. The MiCO Platform uses RabbitMQ for messaging alongside Apache Camel for routing and execution.

MGMs: Depending on the functional and throughput needs of a given installation of AMP, a variety of MGMs may be used. These MGMs may be either automated or manual (i.e. requiring human intervention), and automated MGMs may operate locally on the same server as AMP, in high-performance computing environments, or in commercial cloud services. Because the mechanisms for calling particular MGMs, as well as input and output formats supported by MGMs, may vary, each new MGM will require a small amount of code as an "adapter" to support translation between AMP and the expected inputs and outputs of the MGM.

Due to the volume of data movement potentially required, during this pilot test phase the AMP system will be developed and tested using Linux servers running within Indiana University's internal Intelligent Infrastructure[18] "private cloud" virtual machine and storage hosting environment, with use of cloud-based MGMs and components as appropriate.

Schedule of Activities

The project will span a period of 27 months beginning in October 2018 and ending in December 2020:

Period 1: October 2018-December 2019 (15 months)

October-December 2018

- Consultant and subcontractor agreements will be negotiated and put into place within the first 4-6 weeks, working with IU's Office of Research Administration and Office of Procurement Services.

[18] https://uits.iu.edu/ii

- The first three months of the project will see the core project team beginning to meet monthly. Monthly meetings are important to ensure all partners are in alignment, communication flow is open so that all questions and concerns can be addressed quickly, and to highlight work completed to date. The meetings will continue throughout the length of the project. The only exceptions are months in which project staff meet in person.
- Developer positions will be posted and hired by IU.
- Existing staff at IU and AVP will begin vetting components (including the database, messaging system, and queueing engine) identified in the pilot project.

January 2019 (See Appendix 3 for a detailed agenda)

- The first of three in-person meetings will be held in Bloomington, IN, at IU. All project staff from IU, AVP, UT Austin, NYPL, and Fraunhofer will be in attendance. IU and AVP project staff will lead the meeting.
- The first of three advisory board meetings will be held by phone/online. IU and AVP project staff will lead this meeting.
- Technical planning meetings with Fraunhofer begin.

February-November 2019

- Agile daily scrum and biweekly application demonstration (for project owners) meetings begin. AVP and IU will lead and attend these meetings. Fraunhofer may attend as needed.
- Technical planning meetings, monthly meetings, and ad-hoc developer support with Fraunhofer continue.
- Implementation activities will be undertaken by AVP, IU, and UT. These include:
    - Components compatibility testing and performance testing.
    - Components troubleshooting and configuration.
    - Final selection of components.
- Collections preparation will be undertaken by IU and Collection Partners. These activities include:
    - Gather collection requirements and use cases.
    - Identify target systems and define metadata fields.
    - Define Collection Partner submission packages.
    - Design workflows and test plans for each Collection Partner.
    - Begin preparing media and existing metadata.
- Development to create the following systems occurs:
    - Workflow management system.
    - Storage management system.

December 2019 (See Appendix 3 for a detailed agenda)

- The second of three in-person meetings will be held in Bloomington, IN, at IU. All project staff from IU, AVP, and UT, as well as Collection Partners, will attend. IU and AVP will lead this meeting.
- The second of three advisory board meetings will be held by phone/online. IU and AVP project staff will lead this meeting.

Period 2: January 2020-December 2020 (12 months)

January-October 2020

- Monthly core project team meetings continue through October 2020.
- Collections preparation continues. IU and Collection Partners will:
    - Continue to prepare media and existing metadata for transfer/ingest into AMP.
    - Manually create metadata as necessary.
    - Document workflows.
    - Configure workflows.
- Integration activities begin and are undertaken by IU and AVP. Activities include:
    - The integration of new and off-the-shelf components.
    - The completion of platform architecture.
    - Testing the functionality, performance, and stability of the architecture.
    - The integration of selected MGMs.
- IU and AVP development teams will perform at least three rounds of processing, testing, and review of the generation of metadata for the media and in each Partner's collection. This will involve ingesting media and data sets from each Partner, processing those data sets through workflows containing some (to be defined/as appropriate) number of MGMs, and testing and review of output by metadata specialists.

September-December 2020

- White paper drafting occurs. The white paper cannot be fully completed until the application is functioning and project staff have gathered feedback from the advisory committee and the all-staff in-person meeting.

November 2020

- The third of three advisory board meetings will be held by phone/online. IU and AVP project staff will lead this meeting. Advisors will have an opportunity to see fully functioning AMP and outputs and provide feedback on the application that will be used in the drafting of the white paper.
- The third of three in-person meetings will be held in Bloomington, IN, at IU. All project staff from IU, AVP, and UT, as well as Collection Partners, will attend. This meeting will be used as a demonstration of the system and pilot project results to all project participants. It will also be used as an opportunity to collect feedback on the project and to discuss, prioritize, and document potential next steps.

Expected Outcomes and Benefits

This project can be judged successful if the needs of users, such as those outlined in the user personas, and discovery scenarios, such as that at the beginning of this proposal, can be answered by the metadata created by AMP. The benefits of intensive data processing will include fuller description of the *content* of audiovisual assets rather than description *about* the asset. This content description, in turn, will lead to greater discovery of audiovisual collections, which are

currently, and in general, under-described and underrepresented in our digital heritage collections in the United States.

In order to prioritize allocation of available resources within the pilot project, the full envisioned user experience and range of functionality for AMP envisioned within the previous planning effort will not be an outcome of the pilot. The project team will maintain focus on the core feature set to produce a minimum viable product (MVP). This project will address and demonstrate:

- User authentication
- Use of the workflow manager to configure workflows and connections between MGMs
- Error reporting and handling
- Workflow status tracking and reporting
- Cloud based MGMs
- Local MGMs
- Open-source MGMs
- Commercial MGMs, including setup and tracking of billing costs
- Cost metrics for licensing commercial MGMs
- Automated MGMs
- Human refinement MGMs
- Human generation MGMs
- Classification of MGMs and MGM inputs and outputs
- Use of the messaging component
- Use of the queuing system component
- Use of the logging service component
- Use of the temporary storage component
- Importing existing metadata
- Use of supplementary materials to describe an audiovisual object
- Use of a data warehouse
- Metadata retrieval from AMP via APIs
- Generation of metadata to be used to guide rights, permissions, and access determinations
- Throughput metrics for specific MGMs and the overall system
- Quality metrics on use of varying resolutions of source items for analysis
- Value and utility of AMP metadata in target system

These areas are seen as the essential elements of AMP and will be the points of focus for the AMPPD project. If the AMPPD project is able to demonstrate success based on these elements and thus show that the underlying approach of AMP is valid, it will place the project on sound footing to build out a more complete, functional, easily deployable, production-ready version of AMP for much broader use in a future phase of the project, along with a resourcing and business model for ongoing development and sustainability of the platform.

While limited in breadth and depth compared to the full vision of AMP, this pilot will offer the most robust and promising demonstration of the AMP concept to date. AMP will:

1. Offer an open-source alternative. Whereas the commercial solution is a proprietary platform that supports an ecosystem of proprietary and open-source MGMs, AMP will bring an open-source platform that supports both open and proprietary MGMs.

2. Offer a product that is distinctly well suited for archival content based on two primary reasons:
    a. Archival content is typically described at the collection, series, or folder level; it is atypical for archival collections to be described at the individual item level. This is, in part, an outcome of the widely adopted archival theory of "More Product, Less Process," which calls for archivists to not "continue to let item-level preservation work undermine more rational decisions to arrange a collection only to series or folder level."[19] This project will aim at describing audiovisual assets on a mass scale, in collections where they may have never been described before.
    b. Developers of individual automated MGMs and automated metadata generation platforms often provide demonstrations of their systems using high quality, well produced recordings with the intent of leaving the viewer of the demo with the impression that fully automated solutions are viable. Using these same MGMs or platforms with archival content that is poorly recorded and unproduced will often yield much less convincing results with regard to the ability to utilize a fully automated solution. AMP is different in this regard, using an approach that assumes the need for human metadata generation and refinement in order to achieve quality results. This foundational assumption manifests throughout the AMP architecture, from providing the ability to incorporate manual MGMs into workflows (e.g., manual refinement of an automated speech-to-text conversion MGM, transcription of handwritten text) to manual triggering of MGM task completion and the creation/incorporation of MGMs with user interfaces created for manual entry (e.g., structured forms for the manual review and documentation of donor agreements).

Other key features of AMP that are distinct from other solutions and useful in an archival context are the ability to utilize supplementary materials in support of describing a primary asset, the ability to customize workflows to the needs of specific types of collections, and the ability to flexibly incorporate new MGMs and workflows.

In the AMP planning project white paper, it was mentioned that other non-commercial efforts have fallen short in their ability to demonstrate concepts such as building reflexive workflows from a palette of MGMs, the utilization of machine learning and artificial intelligence, creating quality metadata for lower quality recordings, and the continued cultivation of a metadata warehouse over time. Resource limitations have kept these efforts from delivering robust and compelling results. There is a truly unique opportunity to overcome this pattern through the use of MiCO project code. The alignment of capabilities between AMP and MiCO, with the close timeline between the end of MiCO and the prospective start of the AMP pilot, sets up the AMP

[19] Mark Greene and Dennis Meissner (2005) More Product, Less Process: Revamping Traditional Archival Processing. The American Archivist: Fall/Winter 2005, Vol. 68, No. 2, pp. 208-263. https://doi.org/10.17723/aarc.68.2.c741823776k65863

project to take great advantage of the MiCO project. Moving forward with the AMP pilot project now will seize this unique opportunity to align global funders (European Commission and Mellon Foundation), projects (AMP and MiCO), and project partners (IU, AVP, UT, and Fraunhofer) to advance this effort in a way that has not been possible to date.

There is a true opportunity here to address the next great challenge for audiovisual collections following the need for mass digitization to avoid obsolescence, degradation, and ultimately loss of content. Not only will the work of AMP help to address this challenge, but it will help ensure that past investments that have gone into digitization are not in vain by enabling the generation of metadata that makes this digitized content more widely discoverable, accessible, and usable and informs rights and permissions decisions.

Sustainability of Project Results

As noted above, all source code and documentation developed on the project will be made available as open source in a GitHub repository, and a white paper summarizing project results will be made available through IU's institutional repository.

Metadata successfully created through this pilot project for Indiana University collections will be hosted within IU's digital repository infrastructure, to which the university has made a long-term commitment through the Enterprise Scholarly Systems initiative, a collaboration of the IU Libraries in Bloomington, IUPUI University Library in Indianapolis, and University Information Technology Services, IU's central IT organization. This project is also a component of IU's Media Digitization and Preservation Initiative, the output of which IU, including the Libraries, has made a long-term commitment at the highest levels to preserve and sustain.