Audiovisual Metadata Platform Phase III
Proposed Activities and Rationale

April 4, 2021 (Public version September 23, 2021)

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 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 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 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

[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/

[3] Europeana. http://www.europeana.eu/portal/en/search?q=

[4] DPLA. https://dp.la/search

provided through full-text indexing—without metadata detailing the content of the dynamic files, audiovisual materials cannot be located, used, and ultimately, understood.

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.

Since October 2018, the IU Libraries—in collaboration with the University of Texas at Austin, digital consultant AVP, and NYPL—have worked to help address these challenges through the creation of an open source software platform known as AMP (Audiovisual Metadata Platform),[5] 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 Mellon Foundation, with substantial in-kind staff and computing contributions from IU.

The goal of the still-ongoing current phase of work on AMP, known as AMPPD (AMP Pilot Development), which will be wrapping up in June 2021, has been to develop enough of the AMP system to be able to pilot test it using two AV collections from IU and a third collection from NYPL. The project team has developed a software system that harnesses the Galaxy workflow engine,[6] 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 classification. 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.

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 collections, particularly those that involve significant amounts of spoken word content such as lectures, events, and documentaries, along with oral history interviews and other ethnographic content.

[5] https://go.iu.edu/amppd
[6] Galaxy Community Hub. https://galaxyproject.org/

With these results in mind, we propose 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 AV content, and to use 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.

Ethical Considerations

Testing in this phase of the project is centered on collections from underrepresented cultures and populations, in line with Public Knowledge program goals[7] and as a means to surface and address ethical considerations that surround the use of machine learning tools to extract metadata from audiovisual content. AMP has already encountered these issues in previous phases, resulting in, for example, a limited scope for how facial recognition tools are implemented in the system. After learning about the problems of proprietary facial recognition data sets and surveillance and privacy concerns via sources such as the Algorithmic Justice League, preference has been given to the open source Python-based tool face_recognition,[8] since 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 with the archival and special collections use case where a set of videos can be searched for a known person instead of trying to identify unknown faces using an unknown or proprietary data set.

Working closely with collection managers will be the most direct way the AMP team will have to address ethical considerations in multiple areas: 1) criteria for selection of collections appropriate for use of AMP; 2) human review of both intermediate and final outputs from AMP; 3) access decisions based on both information already known and information provided by AMP; and 4) engagement, where appropriate and possible, of communities represented in and served by collections being processed.. The criteria we use to select collections will be crucial to ensure that we are not exploiting, overexposing, or misrepresenting any materials from historically underrepresented cultures and populations. Those criteria are addressed in the Collection Testing portion of the Work to be Done section below.

Additionally, it is important to note that AMP itself will not provide any access to collections; rather, it provides additional metadata that can then be used as a decision point by a collection manager on providing or restricting access. The outputs of AMP can be pulled into target discovery and access systems, but the results might identify a collection that needs to have access restricted (due to copyright issues or sensitive content identified) or qualified (harmful language identified). The fact that AMP produces metadata does not always indicate that access to that content is going to increase. It does indicate that more information is available for informed decision making.

[7] https://mellon.org/programs/public-knowledge/

[8] https://face-recognition.readthedocs.io/en/latest/readme.html

The AMP team has always had the concern of introducing bias with the machine learning algorithms and training data used by certain 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[9] 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. This has directly resulted in a limited scope for facial recognition as well as close inspection and questioning of results from automated transcription and named entity recognition tools. These choices also follow along with discussion of "refusal to use" as part of creating equitable AI and resisting the idea that the march of technology progress is inevitable, both covered in a recent talk given by Dr. Sarah Myers West from the AI Now Institute.[10]

Ethical considerations also highlight 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 was found to be removing harmful language without our knowledge. The resulting discussion and investigation found that the collection managers want to know that such language is in the transcript and provide a warning to users, but they do not want to remove those words from the transcript due to the need to acknowledge the reality of what was being said on that digitized audio or video recording and the fact that it would not be discoverable if that language was removed. Keeping people as part of the AMP workflow helps ensure consideration for how automated MGM outputs are used. In this next phase of the project we are interested in feedback on these workflow outputs not only from collection managers but from the communities served by and represented in these collections, when appropriate. Such engagement will take place through the collection managers involved.

Involvement of archivists, collection managers, and experts in ethical AI on the project's Advisory Board, discussed below, will also help to keep the project focused on the many dimensions of ethical consideration around the application of AI to cultural heritage collections generally, and specifically to collections involving historically underrepresented populations.

Previous Work

This current proposal is preceded by a 2017 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), was kicked off in late 2018 and continues through June 2021.

AMP Planning Project

[9] https://www.ajl.org/

[10] https://events.iu.edu/siceiub/event/143828-discriminating-systems-gender-race-and-power-in

The 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).[11] 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[12] 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

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

[12] https://avalonmediasystem.org, funded in part by grants from the Andrew W. Mellon Foundation and Institute of Museum and Library Services

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.

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 (for the pilot phase) and Aviary, that offer metadata management and/or discovery related to audiovisual 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

The Audiovisual Metadata Platform Pilot Development (AMPPD) project 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, continues through June 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

[13] 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

automated and human analysis activities, and is being 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.

Using the metadata that exists today, discovery opportunities are 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 had 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 that Wells is (or is not) on a video, but exactly where in the video he appears. And, not only where he appears, but what he says or what is said about him. When ethically applied (see Ethical Considerations section above), this can be a game changer for large AV collections that otherwise have very little description.

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 not music present.

Uncovering the underlying opportunities for metadata capture and exposure of vast audiovisual 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 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. By the end of this project in June, the project team aims to maximize findability and usability of audiovisual 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]

We are well on our way to achieving these goals. By June, the AMPPD project will have produced what was set out for the team to accomplish:

- Architecture

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

    - *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. This includes a fully implemented API.
    - *Ease of us*e: 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. This is achieved through 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.[15] 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.
- MGMs
    - *Appropriate for use*: MGM evaluation includes review of several criteria, including:
        - Accuracy
        - Input formats
        - Output formats
        - Growth rate
        - Processing time
        - Computing resources
        - Ethical considerations
        - Cost
        - Support
        - Integration capabilities
        - Training

Out of scope for the AMPPD project is building a production-ready, out-of-the-box application. As of now, AMP requires skilled IT intervention to set up the environment and tools to enable

[15] https://docs.galaxyproject.org/en/master/dev/schema.html

end users to analyze their collections. Making AMP both easier to use and easier to deploy is a key focus of the work that has been envisioned for the present Phase III funding request.

Related Work

There are two areas to consider for this phase of AMP when describing related work. One area is Artificial Intelligence/Machine Learning (AI/ML) initiatives and the other relates to efforts to create a publishable and deployable system.

AI/ML Initiatives

The *American Archive of Public Broadcasting: From Repository to Resource* grant from the Mellon Foundation to the GBH Archives supports the American Archive of Public Broadcasting's (AAPB) efforts to create a better resource for researchers, educators, academics, and the public. Together with Brandeis University's Lab for Linguistics and Computation, which uses machine learning and artificial intelligence to develop open source tools and workflows, they are capturing detailed metadata from AAPB radio and television programs. This metadata, descriptive information about the people, places, dates and conversations in the archive, is a powerful way to improve access and discoverability of content.

The Brandeis team's tool, called CLAMS (Computational Linguistics Applications for Multimedia Services),[16] is currently being built out to support AAPB needs. CLAMS aims at providing archivists and media researchers with an open platform to access and explore archival audiovisual material to extract insightful metadata, as well as providing computer scientists and developers of content analysis tools with an interoperable platform to integrate their tools for custom workflows and pipelines. Of particular interest to the AMP initiative is the workflow tool that both the CLAMS and AMP platforms utilize. The platform, called Galaxy, [17] is an open source web-based workflow engine for accessible, reproducible, and transparent computational research, originally developed in support of the computational biology community. Both Brandeis and the AMPPD project have repurposed Galaxy for use in managing AI workflows for AV content.[18]

Over the past 18 months, the AMPPD and AAPB/CLAMS teams have met monthly to exchange findings, methodology, and lessons learned, and team members from both GBH and Brandeis sit on the AMPPD project advisory board. In November, members from both teams held a workshop at the Association of Moving Image Archivists (AMIA) conference [19] to share project findings, including machine learning tools and outputs. The workshop—and continued monthly meetings—highlight each project's individual approaches to AI workflows, and also shed a light on areas of synergy.

In the early planning of the AMP project, there was a strong preference for the use of open source technologies for the AMP application. Through testing of AI tools in the AMPPD grant, it became clear that AMP must offer an option for both an open source and a proprietary option for

[16] https://clams.ai

[17] https://galaxyproject.org

[18] More information on the AMPPD project's evaluation of workflow engines and selection of Galaxy is available at https://drive.google.com/file/d/1ZRziZFR3miYgGEebsVqEpB5cn9Fw3rJe/view?usp=sharing

[19] http://www.amiaconference.net/amia-2020-workshops/

each MGM area (e.g., transcription, NER), mainly because the proprietary tools tended to be much more accurate than the open source versions. For that reason, the AMPPD MGM team did extensive research into both types of tools during the evaluation of tools. This enabled AMP to support an ecosystem that joins commercial, proprietary MGMs with open source MGMs (and leaves it up to the user which MGMs to employ). AMPPD has evaluated and/or utilized proprietary services from 3PlayMedia, Amazon Web Services (AWS), Microsoft Azure, Google, IBM, and Mozilla, all of which are active in the ML sphere.

Additionally, many project team members have become involved in the AI4LAM initiative[20] led, in part, by Stanford University. This group evolved from a series of conferences called *Fantastic Futures.* IU staff participated in the second *Fantastic Futures* conference at Stanford in December 2019. Through this loose organization of representatives from across Library, Archives, and Museums communities, individuals meet regularly online and through Slack to share their knowledge about and exploration into AI/ML work, as well as news from the field.

Previous Publishing and Deployment Initiatives

IU has been actively developing and refining their internally-built open source media access tool, Avalon, for nearly a decade. The Avalon Media System is an open source system for managing and providing access to large collections of digital audio and video.[21] IU has been successful in making the application deployable with minimal IT support, and well over a dozen academic institutions have successfully deployed it for their use. Several IU staff who have worked to develop this functionality are also involved in AMP development. This human resources crossover will be invaluable to efficiently build this same deployment functionality into AMP.

In turn, AVP has a long history of building, publishing, and deploying applications of various sizes and complexities. For example, Fixity Pro[22] is a deployable tool that requires little or no IT support to install and run on users' desktop computers (i.e. Windows and Mac-based operating systems). Fixity Pro is used by hundreds of organizations globally, ranging from small organizations such as the Greene County Public Library in Xenia, Ohio and the Maryland Center for History and Culture to large organizations such as the International Federation of Red Cross and Red Crescent Societies and the Mayo Clinic. Additionally, AVP's software-as-a-service (SaaS) platform for media access—called Aviary[23], also available in an open source version[24]—integrates AI tools for transcription with a fully interactive player. Aviary integrates with existing documented APIs from commercial transcription services, e.g., IBM Watson and Trint. All Aviary subscribers get access to IBM Watson and Trint-based transcription for no added cost, however the services themselves do charge for access, and, as a SaaS offering, AVP passes the cost to the subscriber directly. Aviary currently has 118 publishing subscribers representing organizations of all shapes and sizes, including the AIDS Healthcare Foundation, American University, Archives of Appalachia, Council on Library and Information Resources (CLIR), Austin City Limits, Fritz Bauer Institut, George Mason University, Yale University, Louis B. Nunn Center for Oral History at the University of Kentucky, Institute of Southern Jewish Life, Emory University, the National Aquarium, New York University, Oral History

[20] https://ai4lam.org
[21] https://www.avalonmediasystem.org
[22] https://www.weareavp.com/products/fixity-pro/
[23] https://www.aviaryplatform.com
[24] https://github.com/WeAreAVP/aviary-public

Association, Qatar Talking Archives, The Armah Institute of Emotional Justice, Sarah Lawrence University, Women Military Aviators, and the Wisconsin Historical Society, among many others.

Work to be Done

During the AMPPD phase, much of the development effort was devoted to building enough of the platform to enable evaluation of such an approach. Focus was given to evaluating, selecting, and implementing MGMs; the creation of workflows to string together those MGMs; the ability to submit collection content through the workflows; and steps for human intervention (Human MGMs, aka HMGMs). The project team focused on areas of application development that supported the essential goal of enabling the evaluation of the platform, and some areas that would be necessary in a production setting were left for later.

In this next phase we will focus on the following areas to bring the pilot application to a place in its development cycle that makes it production-ready:

System Robustness and Resilience

There are a few areas where we will do additional work in Phase III to ensure a robust and resilient system:

1. Role-based access control – During the pilot, access to the AMP application was limited to team members, so we considered implementing a role-based access control module a low-priority task. In a production setting, though, robust access control is paramount to ensure appropriate permissions are granted to users depending on their role and needs (see Appendix 1), and security and privacy requirements for collections are supported by the system.

2. Human Intervention – One of the unique features of AMP is the integration in the workflow of metadata generation mechanisms of the ability for human intervention, i.e., the ability to have a human review and correct automated results before those results are used in other steps of the workflow. Unlike automated steps, human intervention steps may take several days or more to complete depending on availability of staffing. The Galaxy workflow engine used by AMP was not totally prepared for workflow steps that can take several days to complete, and so we had to customize it to ensure that pending human-action steps do not lock computing resources while waiting for human intervention.

   One aspect that remains to be addressed is the scalability of this feature. We may need to consider the tuning of the configuration with respect to the number of job runners and workers.

3. Human Intervention Tools – AMP currently integrates a transcript editor and an editor for refining extracted named entities. Both need some work to become production-ready:
   a. We have selected the open source BBC Transcript Editor for collection staff to use when correcting automated transcripts. This tool is powerful in many ways, but it never left the labs at BBC, and important functionality is missing such as the

ability to edit timestamps. We have also run into a few hindering bugs. While we have addressed the most important ones, others remain to be worked on in Phase III, along with a short list of features that our collection staff performing transcript review have requested, such as the ability to use diacritics and italics.

   b. The NER output editor was adapted from the Avalon Timeliner tool.[25] Currently, this tool shows an audio player only, given its original use case. For AMP Phase III, we plan to add a video player to improve user experience and functionality.

4. Task Management tool - AMP offers hooks for integration with different task management systems, used to manage and track human intervention steps. Currently the only tool integrated with AMP using those hooks is Atlassian Jira. Besides this proprietary tool, we intend to, in Phase III, add a free or open source option to enable AMP implementations by institutions who may not desire to pay for a tool. Based on our knowledge of task management tools commonly used in libraries and archives, our current plan is to develop an additional integration with Trello,[26] which is not open source, but which offers a popular free tier.

Packaging, Deployment, and Technical Documentation

While AMP code is currently available for download via GitHub, the code currently is not packaged for easy deployment, and deploying it requires some manual steps tailored to the particular environment in which it is being installed. To ensure ease of deployment of the multiple instances required to support the lifecycle of an application, but also the deployment of AMP by other institutions who may not have IT staff at all or IT staff sufficiently familiar with the entirety of the technology stack adopted by AMP, we will utilize a multi-tier approach to creation of deployment methods, with the end goal of providing a containerized environment for AMP. By splitting the packaging into multiple tiers, institutions can jump on board at the level they feel comfortable and deploy AMP in a way that best suits their needs.

Containers, as supported by Docker[27] and other systems, have become the technology of choice for packaging applications, providing portability between operating system platforms, including cloud services such as Amazon Web Services, Google Cloud Platform, and Microsoft Azure. This technology simplifies the tasks of installing, managing, operating, upgrading, and uninstalling complex software because it bundles together the application and its dependencies.

The first tier is to install AMP directly onto a specific operating system. We will provide detailed documentation and scripts to automate the installation and configuration process and test this process in order to refine and improve the documentation. These scripts will be used by the tiers specified below, but can also be used by institutions with special requirements who are comfortable with supporting this level of integration.

Building on the OS-level installation scripts, the second tier is to create the scripts and configuration required to generate deployable container images for the different components of AMP. These images will be specified by a Dockerfile (or equivalent) to produce repeatable

[25] https://timeliner.dlib.indiana.edu
[26] https://trello.com
[27] https://www.docker.com

images. The different component images can be distributed to image repositories, such as Docker Hub for deployment into container runtime systems.

After the container images are created, we will create scripts and configuration to provide the orchestration required to bring up the different containers into a coherent AMP service, providing the third tier. A possible choice is using Helm Charts to orchestrate deployments with Kubernetes,[28] but Terraform[29] and other technologies will also be explored. We envision this tier as the primary installation method for nearly all users, whether using local or cloud computing environments, so creation of detailed deployment documentation will be essential.

In addition to these needs, AMP currently lacks a scripted way to perform database migrations during deployments. It is the industry standard to provide tooling for this important deployment activity, and we plan to address this need in Phase III of the AMP project as part of our packaging and deployment work.

User Experience Evaluation, User Interface Development, and User Documentation

As mentioned before, the focus of the AMPPD phase of the project has been enabling the evaluation of the platform as a viable solution for the problem it proposes to address. This inevitably implied moving to the back burner the implementation of friendly and intuitive user interfaces for some of the supporting activity that needs to occur within the application.

In Phase III we propose to address the following important areas:

1. Creation of workflows: AMP currently uses the Galaxy interface for workflow creation and editing. However, our goal is to not need to send users to the Galaxy interface for anything; for consistency and ease-of-use, they should be able to perform all their tasks from the AMP interface. Implementing an interface for workflow creation in AMP that adheres to AMP look-and-feel standards, and which is consistent with how AMP implements other activities, will reduce significantly what collection managers need to learn to work with AMP.

2. Manipulation of data in AMP: There are a variety of issues related to data manipulation:
    a. The only way to upload AV content in AMP is via a batch process. Updating that data is also only done using the batch process.
    b. There is no mechanism in the AMP user interface to delete items, so they must currently be deleted by the developers, when necessary
    c. There is no way for a collection manager to navigate their collections.

    User Interface pages for these activities have been designed, but have yet to be implemented (see Appendix 2).

3. Intermediary files: We started AMP with the concept of processing the primary AV files that comprised the test collections. It did not take us long to realize that we would very often also need to process intermediary results. For instance, if a workflow generates a

[28] https://kubernetes.io
[29] https://www.terraform.io

transcript for a video in a collection, and the collection manager later realizes that it would be good to have a WebVTT[30] file generated from that transcript, currently they would either need to go to Galaxy to do that or resubmit the video file to a workflow that again generates a transcript and then uses the transcript to generate the WebVTT file.

We have already investigated the path to solving this in AMP and plan to implement it in Phase III.

4. Workflow Administrator pages: AMP uses the Galaxy workflow engine (see Project Technology). Even though Galaxy is robust and offers a large number of different controls and APIs, its user interface (UI) for managing jobs is lacking. For Phase III, using the rich offering of Galaxy APIs, we will implement additional functionality in the AMP UI to facilitate this task for workflow status and administration.

The Product Owner will develop and improve user-focused documentation for AMP, building on the existing AMP User Guide[31]—to complement the technical documentation discussed above—based on feedback from collection managers and other users.

In addition to new feature development, we plan to carry out user experience work to evaluate currently existing functionality with collection managers and conduct user research to inform and evaluate design of new functionality to be implemented in Phase III. Research methods will include surveys, user interviews, and usability testing of mockups as well as finished products.

Subjects for UX research will include both collection managers and catalogers/metadata specialists at IU along with staff from NYPL.

MGM Evaluation Interface

In the AMPPD project, we tested accuracy for over 20 proprietary and open source tools in eight MGM categories over samples from our three partner archives. An early lesson we learned was that any accuracy benchmarks promoted by tool providers are generous at best; the tools are designed for particular use cases and likely not trained on the types of archival footage that AMP users will be analyzing. We found that accuracy can vary widely depending on many different variables, such as audio/video quality, language, speaker dialect, video color, background noise, and content, to name a few. We made our selections of MGMs for inclusion into AMP based on our limited samples and use cases, but this should not be interpreted as endorsement. We recommend that users test AMP MGMs on samples from their own collections and set their own benchmarks to inform appropriate MGM selection and to manage expectations and understand the potential human remediation necessary for "acceptable" output.

To test accuracy for the AMP pilot, we created ground truth data for each sample for each tool tested, normalized output from each tool tested, and wrote Python scripts to quantitatively compare ground truth against tool output. We also converted those outputs into data formats

[30] Web Video Text Tracks, a World Wide Web Consortium standard format for time-based text. https://www.w3.org/TR/webvtt1/

[31] https://wiki.dlib.indiana.edu/display/AMP/AMP+User+Guide

useful for human visualization (e.g., CSV for viewing in Google Sheets, tab-delimited text files to view labeled timestamps alongside audio in Audacity), so collection managers could qualitatively assess MGM outputs and better understand the frequency and nature of inaccuracies. While we have been compiling these scripts in a public GitHub repository for use by any future AMP users, the overhead in implementing such a testing framework will prove too high for many.

We believe that implementing such testing scripts as an evaluation interface within the AMP platform will empower users to more easily predict the effectiveness of MGMs for their particular collections and to calculate the risk involved in applying artificial intelligence to their use cases.

We envision that for each MGM, a user will be able to submit ground truth datasets for their collection samples (some of which can be created through the use of existing human MGMs), send their samples through the MGM, run accuracy testing against ground truth, and view results in the interface. Users should also be able to visualize and compare ground truth and MGM output for several MGMs or at least be able to export data in a format that can be visualized by a spreadsheet tool or other tools.

We also plan to include guidance within the interface on applying additional evaluation criteria we created during the pilot (e.g., cost, processing time, ethical considerations) to help users select the best MGM for their use cases.

The evaluation interface will consist of:

1. Testing interface: Templates and guidance will be provided in the interface for creating ground truth data in spreadsheets or other easily accessible tools. Users will then upload ground truth data through the interface to be converted to the AMP JSON format, run accuracy tests, and view scores relevant to the MGM, such as accuracy, precision, recall, and word error rate.
2. Visualization interface: MGM and ground truth results will be available to compare side-by-side in a tabular format and available for download for visualization in external tools.
3. Evaluation criteria guidance: A page for general evaluation criteria and specific MGM guidance will help users assess MGMs for their own use.

Collection Testing

Once the work described above has been completed, we plan to test this new version of AMP, with a focus on collections selected based on inclusion of materials from historically underrepresented cultures and populations. At IU, materials for testing will be gathered from across various larger collections, such as the University Archives, IU Bicentennial Oral History Project,[32] Center for Documentary Research and Practice,[33] and Archives of Traditional Music,[34]

[32] https://200.iu.edu/signature-projects/oral-history/index.html

[33] https://cdrp.mediaschool.indiana.edu

[34] https://libraries.indiana.edu/archives-traditional-music

each of which has significant collections or collection subsets focused on the experience of people of color and other historically underrepresented communities. Other collections to be considered that are specifically centered on such communities include the Archives of African American Music and Culture[35] and Black Film Center/Archive,[36] which focus on the experiences of African Americans and the Black Community.

Criteria for selecting collections for testing will include the following:

- Approximately 50-100 hours of content
- Content is about and/or created by historically underrepresented cultures and populations
- Collection managers are able to establish which automated content analysis is allowed (based on documented donor agreement on file or consultation)
- Where possible and appropriate, communities represented by these collections are consulted for feedback on metadata outputs from AMP analysis and resulting access decisions

Current AMPPD partner NYPL will also continue as a partner in Phase III. NYPL has an extremely diverse range of AV collections to choose from, and based on discussions to date, NYPL will focus their testing on two collections from the Schomburg Center for Research in Black Culture:

> **Communications Excellence to Black Audiences (CEBA) audio and moving image collection**
>
> The Communications Excellence to Black Audiences (CEBA) audio and moving image collection consists of 1,112 audio items and 2,750 moving image items. The World Institute of Black Communications (WIBC) produced the CEBA awards from 1978 to 1991 to recognize programming created for Black people. A recognized precursor to the NAACP Image Awards, the collection is the award entries submitted by advertising, radio, and television entities from across the United States. When acquired in 1986, Howard Dodson, the Director of the Schomburg Center for Research in Black Culture said, “This donation ranks among the most significant in the Schomburg Center’s [then] 60-year history” as a record of Black creatives and the enormous changes of the Black image in the media.
>
> Audio recordings include product advertisements, public service announcements, and political campaign announcements. Moving image recordings include product advertisements, public service messages, public affairs programming, news productions, dramatic productions, documentary programs, and music videos.
>
> Television award nominees represented in the collection include (but are not limited to) Positively Black, Black Nouveau, Our Voices, BET News, Video Music Box, South Africa Now, The Oprah Winfrey Show, A Different World, Eyes on the Prize, Teen Summit, The Women of Brewster Place, Screen Scene, Best Talk, Essence: The

[35] https://aaamc.indiana.edu

[36] https://bfca.sitehost.iu.edu/home/

Television Program, Visions, Insights, City Line, People Are Talking, Faces, and Common Ground.

All-American News moving image collection
The All-American News moving image collection consists of 39 newsreels comprised of more than two hundred and fifty segments produced for Black audiences from 1944-1946. Supported by the US government and the Office of War Information (OWI), Black filmmaker William Alexander filmed all over the world highlighting Black contributions to World War II. He co-wrote with veteran journalist Claude Barnett and other members of the "Black brain trust" hired by the OWI to communicate with Black audiences in segregated movie theaters. As the tagline spelled out: "All-American News brings you our people's contributions to America and freedom." Alexander and his colleagues' newsreels supported the US war effort just as they undoubtedly thought of it as a tool for the Black media's campaign for Double Victory – victory over fascism abroad and racism at home.

Distributed by veteran Hollywood exhibitor Emmanuel Gluckman, All-American News reached Black moviegoers in more than 900 theaters across the country from 1942 until the end of the war. From the war updates, featuring footage of the all-Black 92nd Infantry, to sports, music, literature and human-interest stories, the shorts reflected Black life as opposed to distorting it through accepted stereotypes used to promote and justify segregation. Documentary in nature, the shorts have been described as "some of the best and first consciously produced, positive images of Black people other than race movies."

While collection managers will be the primary contact points for the media being used and the direct evaluation of AMP, the AMP project team will coordinate with collection managers when possible and appropriate to engage with community groups or members of historically underrepresented cultures and populations for participation, evaluation, and feedback, both of the treatment of materials and of the metadata generated for furthering discoverability and access.

The goal for this phase will be to provide AMP as a service through which collection managers can select collections and sets of items, load them for processing, select the desired workflows to produce the metadata they would like to see (transcripts, named entity recognition terms list, facial recognition, segmentation, scene/shot detection), run those workflows and evaluate the outputs, and use those outputs with selected applications and services for testing. This will include testing output with access platforms including Avalon Media System and Aviary.

Project Meetings and Reporting

Because of COVID-19, it is extremely difficult to know exactly how to approach planning for in-person project meetings. In past projects, the AMP team has met at important milestones (e.g., project kickoff, midpoint, and wrap-up). We are taking a conservative approach and assuming it will be unlikely that an in-person meeting will be possible in 2021. At the same time, we are taking a hopeful stance that one in-person project meeting will take place at the end of 2022.

For 2021, all project meetings will be held online via Zoom, or a similar web-based meeting platform. Meetings are prepared and led by a project manager, although the core team plays an important role in preparing agendas. Several different standing meetings will take place on a weekly, biweekly, and/or monthly schedule: core team, MGM team, collections team, and an all-team meeting. The development team will continue to take an Agile approach with two-week sprints. Typical meetings that support this approach (e.g., daily standups, biweekly sprint planning and backlog grooming) will also be an important part of maintaining cross-team communications and project momentum.

- The kickoff meeting in Fall 2021 will be used to ensure that all project participants are in alignment with regard to project scope, process, responsibilities, roles, timeline, and deliverables. This meeting will also serve as an opportunity to gather momentum and support team building through online face-to-face interaction. All project participants will participate in this meeting.

- The second project meeting will take place in Summer 2022. This meeting will include all project participants and will serve as an opportunity to regroup as a team and ensure alignment, and plan for the next half of the project.

- The final wrap-up meeting in late 2022 will be an opportunity to meet in person in Bloomington, Indiana, at IU. It will focus on demonstrations of the production-ready system and other deliverables 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.

Resource Needs

The IU Libraries plan to devote significant existing staff resources to project direction and software development management and also to contribute staff resources in the areas of collections expertise, metadata analysis, and IT systems engineering. The IU Libraries will also contribute both local computing resources and cloud computing service resources needed by the project. However, additional resources are requested from the Foundation to support a number of needed functions, including high-level project management, software development, subject-matter expertise (SME), and collections staff engagement. 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 one senior software developer to lead continued design, coding, and testing of the AMP system
    - Contracted software development services to contribute to design, coding, and testing of the AMP system
    - Consulting services from AVP to support AV and MGM subject matter expertise and project management
    - Labor costs for the participation of NYPL as an external Collection Partner

- Travel and meetings:
    - Travel and hosting costs to support one project meeting and promotion of AMP project work at relevant conferences

Dissemination

New versions of the AMP application will continue to be released under an Apache 2.0 open source license, with source code made available via GitHub. In addition, the application will be made available as a set of container images for deployment on local computing infrastructure or in cloud computing environments such as Amazon Web Services, Google Cloud Platform, and Microsoft Azure, via technologies such as Docker, Kubernetes, and Terraform.

AMP project staff will report on project results 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, Fantastic Futures, and others.

Project Organization

The project will be led by the 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 the current Mellon-funded AMPPD project and previous 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 three project team meetings to review status and conduct near‑term and strategic planning.

AVP will work on the project by providing project management, expert advice and analysis on workflow design, metadata implementation, and choice/implementation of MGMs.

A diverse advisory board of six to eight members will be established for the 18-month duration of the project to provide advice in the areas of AI and machine learning bias; system architecture and design; application of machine learning-based MGMs for audio, video, and natural language processing; user needs of archivists and collection managers; community member engagement; metadata; and other aspects of AMP system development and potential use. The advisory board will meet by phone or videoconference at least two times over the course of the project. Members will be invited from amongst the current AMPPD advisory board, supplemented with invited experts from additional areas.

Current AMPPD Advisory Board:

- Gerrit Bruns, Research Assistant, German National Library of Science and Technology
- Michael Giarlo, Technical Manager, Stanford University Library

- Caitlin Hunter, Head of Recorded Sound Section, Library of Congress
- Casey Davis Kaufman, Associate Director, WGBH Media Library and Archives
- Brian McFee, Assistant Professor of Music Technology and Data Science, New York University
- James Pustejovsky, Feldberg Chair of Computer Science, Brandeis University
- Dirk Van Dall, VP, Advanced Research, BAMTECH/Disney Streaming

IU-Funded 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 board. He has over 25 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 current AMPPD project, IMLS-funded Variations3 project, IMLS and Mellon-funded Avalon Media System project, and Mellon-funded Integrating Licensed Library Resources with Sakai project. He has also served as a member and chair of the Steering Committee for the Samvera open source digital repository software community.

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 and serve as liaison to the User Experience Specialist for work on UX evaluation and UI design. 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 Product Owner on the AMPPD project since October 2018 and 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 as Product Owner for the university's Human Resources Management System.

Brian Wheeler, Senior Systems Engineer (.15 FTE), will serve as a technical consultant on systems architecture and implementation and will lead packaging and deployment work on the project. 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.

Julie Hardesty, Metadata Analyst (.15 FTE), will be the primary interface between the AMP project team and Collection Partners, helping to select appropriate test collections and workflows, and assisting collections with testing workflows and evaluating MGM outputs. 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 has worked with metadata for audiovisual materials on projects including Avalon Media System, HydraDAM2, and the IU Media Digitization and Preservation Initiative. She has also been active in metadata activities and product development within the Samvera Community, currently serving as Product Owner for Samvera's Hyrax digital repository toolkit.

Jon Cameron, Digital Media Service Manager (.15 FTE), will work with Collection Partners to setup and utilize the MGM Evaluation Interface and facilitate the 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 (.15 FTE), will contribute to the design of workflows for the selected test collections, with a particular focus on the use of Human MGMs, working closely with Julie Hardesty, Jon Cameron, and the software development team. 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.

Isuru DeSilva, IT Project Manager (.20 FTE), will serve as Scrum Master for ongoing AMP development. He has worked for the IU Libraries since 2018 as Scrum Master on multiple projects, including new text, image, and research data repository platforms based on Samvera Hyrax, and IU's new ArcLight-based Archives Online platform for archival collections discovery.

A **User Experience (UX) Specialist** (averaging approximately .05 FTE over the course of the project) in the Enterprise Systems division of IU's University Information Technology Services organization will be responsible for work on developing recommendations and specifications for user interface designs and improvements, based on user testing work performed by AVP staff. The UX Specialist will be engaged at specific points in the project when this work is required. For example, they will perform design work when new web interfaces are ready to be built.

Grant-Funded Staff

Ying Feng, Senior Software Engineer (1.0 FTE) will be paid using grant funding to lead the technical aspects of AMP development, including feature design, coding, and testing. She has served in this role for the AMPPD project since early 2019. Ying has 15 years experience in software development, especially with large scale open source web applications. Prior to joining the AMPPD project, she worked as a senior developer on the Avalon Media System project for over a year, and the Kuali Financial System project for over 10 years.

A contractor **Software Engineer** (1.0 FTE) will be hired by IU using grant funding for the 18-month duration of the project through an existing IU contract staffing vendor, to assist in 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 Human 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 also contributed development leadership and services 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. 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 and metadata subject matter expert (SME). 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.

Shawn Averkamp, a Senior Consultant with AVP, will provide subject-matter expertise (SME) and technical support in the design of the evaluation interface. Her work at AVP has focused on exploring creative, user-focused solutions to both novel and common data challenges, from leading experimentation and assessment of machine learning tools for audiovisual metadata, to supporting supply chain analysis of cattle transactions and deforestation in Brazil, to mapping large-scale data migrations. Before joining AVP, she supported all stages of the data lifecycle in many different areas including metadata management, data modeling, digital humanities projects, crowdsourcing platform development, and data curation. Since joining AVP in 2019, Shawn has consulted with the National Wildlife Federation, Library of Congress, Smith College, Indiana University, and Denver Public Library.

Bertram Lyons, Managing Director for Software, and Partner at AVP, will provide technical guidance and implementation vision for all areas of the project. He leads AVP's development team to create innovative, flexible, and user-center software for AVP projects and clients. His background specializations include the acquisition, management, and preservation of documentary, research, and cultural heritage collections.

An AVP consultant to be determined will conduct user testing of the AMP interface to inform user interface and interaction design work carried out by the UX Specialist at IU and implemented by the development team.

Subgrantees

New York Public Library will be engaged as a subgrantee on the project under the direction of **Greg Cram**, Director of Copyright, Permissions and Information Policy, with grant funding used to support the time of Schomburg Center curator **Shola Lynch**, metadata archivist **Alex Duryee**, and metadata specialist **Sarah Rubinow** in using the AMP platform to design, execute, and evaluate workflows for NYPL's selected test collection(s). In addition, **Jay Haque**, Director of Infrastructure and Operations, will provide feedback on potential system deployment models and coordinate NYPL testing of containerized deployment packages for AMP.

Project Technology

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.

AMP's current technical architecture is shown in Figure 1:

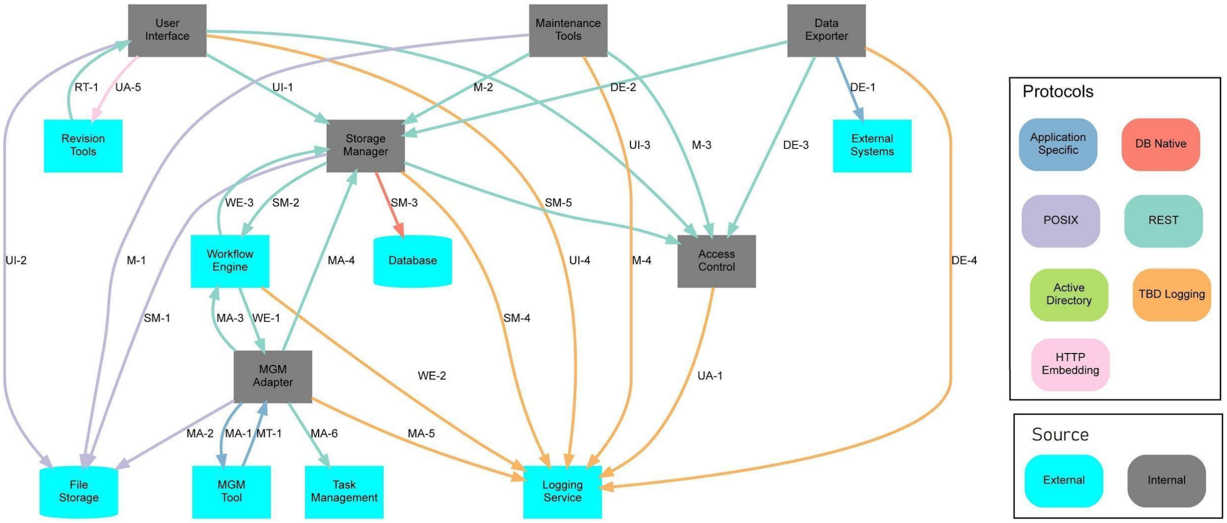

Figure 1 - AMP System Architecture Diagram

The AMP architecture involves a combination of existing software components and new components developed by the AMP development team.

Workflow Engine: The Galaxy[37] open source, web-based workflow engine is a robust application that provides all the features one expects of workflow engines, among them: ability to create, update, delete workflows; ability to add new custom tools to be used in workflow steps; ability to run workflows and resume paused workflows; ability to tell when a job was run, how long it took, which input files were used, 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.

Another important Galaxy feature is its ability to do file type checking during workflow creation to prevent the creation of malformed workflows. This is an excellent feature, but the file types native to Galaxy are not refined enough for our purposes; for instance, a binary file is a binary file, whether media or not. 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 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 to AMP users.

Programming language: The AMP project uses Java for the development of its core functionality on the backend, with Vue.js as the JavaScript framework for frontend needs.

[37] https://usegalaxy.org/

Java—a class-based, object-oriented programming language — was selected for various reasons: on the technical side, it is portable and architecture-neutral, robust, and secure. In addition, there is a good offering of tools to improve Java developer productivity. Java is still quite popular for enterprise application development,[38] but the architecture of AMP would allow for the backend core to be rewritten if needed in the future, while still retaining existing Vue.js/JavaScript/HTML code for the frontend and the Python-based Galaxy workflow engine.

Galaxy itself is developed in Python, and Python is used in the creation of MGM adapters and of locally-developed MGMs, although Galaxy itself does not require adapters to be written in this language; as long as the code is converted to an executable, it is acceptable by the workflow engine. The revision tools selected for the project happen to be written using the React.js framework and are easily integrated with the Java core codebase using packaging tools.

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. The team chose the relational database PostgreSQL based on its robust support for blobs and JSON, while at the same time offering the advantages of the relational model for queries.

Storage Infrastructure: Because of the large size of audio and video media files, a design goal of the system was to minimize as much as possible network movement and copying of AV data. The system supports submission of content available on a filesystem shared with the machine on which the AMP platform components and local MGMs are running. Individual MGM steps may generate intermediate transformations of metadata that require temporary storage. The implementation of the UI that will enable ingestion of content from a user's local machine will be part of the work on Phase III.

Storage Manager: Storage and retrieval functions for content and metadata are mediated through a storage management module of the core AMP codebase 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.

MGMs: Depending on the functional and throughput needs of a given installation of AMP, a variety of metadata generation mechanisms (MGMs) may be used. These MGMs may be either automated or manual (i.e. requiring human intervention [HMGMs]), 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 the pilot phase the AMP system was developed and tested using Linux servers running within Indiana University's internal Intelligent Infrastructure "private cloud" virtual machine and storage hosting environment, with use of cloud-based MGMs and components as appropriate.

[38] See PYPL PopularitY of Programming Language index. https://pypl.github.io/PYPL.html

Task Manager: As the workflows get to manual steps (HMGMs) of the workflow, AMP creates tasks for the humans performing the manual work. During the pilot, the decision was to not create yet another task management tool when there are hundreds of offerings in this domain. Using APIs, AMP creates tasks in the task manager component and can also close those tasks when AMP detects the task has been completed. We implemented this strategy using Atlassian Jira as this was the task management tool known to our staff performing the manual steps during AMPPD; in Phase III we intend to implement a free or open source alternative.

Revision Tools: Currently AMP has integrated two revision tools: to revise automatically generated transcripts we chose the open source BBC Transcript Editor[39] for its features and ease of adoption. To revise the automatically extracted list of named entities, we use the Avalon Timeliner, a tool that already deals with AV content and point-in-time annotations. We adapted it to AMP's needs by removing the bubble annotations and keeping the point-in-time annotation markers. There is the potential of using Avalon's Structural Metadata Editor to revise the output of segmentation tools due to its ease of use, editing features, and the waveform display that serves as a visual aid in identifying segments. Other revision tools can be investigated for other types of output as the need arises.

High-Performance Computing (HPC) environment: By using adapters around tools in Galaxy, AMP users can mix and match, in the same workflow, MGMs that are executed locally in the AMP server, in the cloud, or in IU's HPC environment. AMP is currently set up at IU to take advantage of GPU-capable HPC resources for Kaldi and the INA Speech Segmenter.

Data Exporter: AMP's data exporter is a set of APIs that allow target systems to request AMP deliverables for an item or a collection of items; a push option is planned. The deliverables are packaged in a JSON formatted file; when a deliverable is in itself a file, AMP provides in the JSON a URI referencing the Web-accessible file.

Schedule of Major Activities

The project will span a period of 18 months beginning in July 2021 and ending in December 2022:

Period 1: July-October 2021 (Project Kickoff and Ramp Up)

- Consultant and subcontractor agreements will be negotiated and put into place within the first four weeks by the PI, working with IU's Office of Research Administration and Office of Procurement Services.
- The first three months of the project will see the core project team beginning to meet twice monthly, with meetings led by the Project Manager (PM). These meetings are important to ensure all partners are in alignment, communication flow is open so 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 the Project Alignment meetings take place (e.g., Kickoff, Mid Point, Wrap Up).

[39] https://github.com/bbc/react-transcript-editor

- Software development work, Agile daily scrum, and biweekly application demonstration (for any/all stakeholders) meetings begin and are led by the Scrum Master and Product Owner. During the first 4 months, software development will focus on these areas:
    a. Define strategy of the integration AMP/Galaxy for role-based access control.
    b. Implement the already designed application pages for Collection creation and content navigation.
    c. Create a first iteration of the application packaging
    d. Define details for the deployment offerings
    e. Write user stories for the MGM Evaluation Interface
    f. Start engagement with UX specialist on the design of remaining UI pages
    g. Add a video player to the NER Revision Tool
    h. Implement solution for using Intermediary files as input to MGM workflows
- The first of three all-hands Project Alignment meetings will be held online in September or October. All project staff from all teams, including collection partners, will attend. The meeting will be organized by the PM and PI.
    - Objectives: Gain team alignment on project history, goals, and objectives; Develop team camaraderie through team-building activities; define and approve project roadmap and timeline; define communication processes.

Period 2: November 2021-July 2022

- Software development work, Agile daily scrum and biweekly application demonstration (for any/all stakeholders) meetings continue and are led by the Scrum Master and Product Owner (PO). For these next 9 months, the software development goals are to:
    a. Finalize initial UX design
    b. Complete the AMP frontend development, implementing two of the remaining pieces for full integration with Galaxy: access control and workflow creation. To provide for user independence, priority will be given to UI pages that support user-driven tasks that currently depend on development staff to be completed.
    c. Work on improvements to the BBC Transcript Editor
    d. Work on one or more iterations of the application packaging and deployment solution
    e. Solidify and implement most of the deployment offerings
    f. Complete the MGM Evaluation Interface
    g. Address any scalability issues with the Human MGMs
    h. Conduct testing
    i. Address new needs that emerge from the Collection Testing effort
    j. Implement new MGMs as requested by Collections staff
    k. Redesign and enrich the AMP User Guide
    l. Update technical documentation as needed
- Twice monthly technical and core team planning meetings, monthly all-team meetings, and ad-hoc developer support (as needed) continue and are led by the PM and PI.
- IU and NYPL will select collections, gather files, and select workflows, in consultation with the AMP project team.
- IU collection partners will use IU's production installation of AMP to run workflows and evaluate MGM outputs.

- NYPL will begin to run workflows and evaluate MGM output using an IU-hosted installation of AMP.
- Initial work on packaging and deployment will be completed by the software development team.
- User testing of workflow creation, data manipulation, and administrator interfaces will be conducted, led by AVP
- First iteration of end user documentation is written in summer 2022 by the Project Assistant in collaboration with assistance from the PO.
- The second of three all-hands Project Alignment meetings will be held online in June 2022. All project staff except collection partners will attend. The meetings will be organized and led by the PM and PI.
    - Objectives: Work through any communication process concerns; maintain team camaraderie through team-building activities; review and revise roadmap and timeline as appropropriate; demonstrate development successes; discuss and problem-solve development challenges; gain team alignment on any outstanding issues.

Period 3: August 2022-December 2022

- Software development, Agile daily scrum and biweekly application demonstration (for any/all stakeholders) meetings continue led by Scrum Master and PO. During the final five months, software work will include the following:
    a. Implement changes to address recommendations from the UX Testing
    b. Export AMP deliverables to target systems such as Avalon and Aviary
    c. Engage with partners for testing and validation of application packaging and different deployment strategies
    d. Implement the Workflow Administration page(s)
    e. Finalize deployment documentation (built alongside the packaging and deployment work)
    f. Integrate AMP with Trello as a free option for task management
- PO gathers feedback on end user documentation from collection managers and other users. Project Assistant, with assistance from the PO, completes the documentation, based on feedback and covering additional application features.
- Twice monthly technical and core team planning meetings, monthly all-team meetings, and ad-hoc developer support (as needed) continue and are led by PM and PI.
- External collection partners will test installation and use of AMP container images and packaging in cloud or local server infrastructure, supported by the Senior Systems Engineer.
- User testing of MGM Evaluation Interface will be conducted, led by AVP, and improvements designed by the UX Specialist will be made based on this testing.
- Partners will use MGM Evaluation to test MGMs and refine workflows.
- Final release of packaged AMP system, along with documentation, is made available for download and installation by the development team, led by the PO.
- The third of three all-hands Project Alignment meetings will be held in person in Bloomington, IN, at Indiana University in December 2022. All project staff, including collections partners, will attend.

- Objectives: Demonstrate the system and 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 Outcome and Benefits

At the end of AMP phase III, we expect to have a robust software system that can be used by libraries and archives to design, test, and execute custom metadata generation workflows that are appropriate for their AV collections, along with documentation of the system and of the project team's experiences in using AMP to help produce metadata for a diverse set of collections from IU and NYPL. The system will be packaged so that it can be deployed by institutions with a range of levels of technical staff and expertise or by service providers that could explore offering AMP in a SaaS model. The AMP system will be promoted and demonstrated at conferences focused on a research library and AV archive audience, and the project's wiki and Github presence will both be designed to serve as starting points for exploration of the platform, with links to user and technical documentation, demonstration videos, and a form to request limited-time access to a test instance of the system hosted by IU for evaluation purposes, similar to the model under which IU has long made Avalon Media System available for evaluation.[40]

In addition, the MGM Evaluation Interface tool and its accompanying documentation will allow librarians, archivists, and other managers of AV collections to test and evaluate machine learning-based and other automated tools against ground truth data from their particular collections and use cases, in order to evaluate risk vs. benefit in utilizing various open source and commercially available tools. Currently no such tool is available in the market.

The metadata produced during the testing work of phase III will support increased discoverability of and (as appropriate based on ethical and rights concerns) access to AV collections from IU and NYPL, with a focus on collections involving materials from historically underrepresented cultures and populations.

Ultimately, development, production deployment, adoption, and use of AMP will lead to greater discoverability of audiovisual collections, which are currently, and in general, underdescribed and underrepresented in our digital heritage collections in the United States. 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 continued 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 access, rights, and permissions decisions.

Project Sustainability

All source code and documentation developed on the project will be made freely available as open source in a GitHub repository, and the IU Libraries, through its dedicated Digital Media Software Development team of 5 FTE, intends to continue to maintain and support AMP for at

[40] https://www.avalonmediasystem.org/try-out-avalon

least five years following the conclusion of the grant. However, the real key to sustainability is to create something so valued that people and institutions insist on sustaining it, and broad community engagement and adoption are the most important means of ensuring the ongoing availability and maintenance of any open source system. To that end, IU, in the course of this grant and following it, will actively demonstrate and promote the tool for broad use by collecting institutions and potential SaaS service providers and will seek additional code contributions, tool integrations, and revenue-generating consulting engagements from adopters of the system.

Metadata successfully created through this phase of the 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.

Additionally, at any point during the project, participating institutions will have the opportunity to export and store their metadata on their local storage infrastructures, as they wish.