IMLS Research Grant: June 1, 2015
Improving Access to Time-Based Media through Crowdsourcing and Machine Learning
WGBH Educational Foundation, American Archive of Public Broadcasting and Pop Up Archive
Abstract

WGBH Educational Foundation (WGBH), in partnership with Pop Up Archive, is seeking a Research Grant to address the challenges faced by many libraries and archives trying to provide better access to audiovisual collections through online discoverability. This 30-month project will combine technological and social approaches for metadata creation by leveraging scalable computation and engaging the public to improve access through crowdsourcing games for time-based media. Given the time-based nature and increasing scale of audiovisual collections, it is difficult for this content to be made accessible by standard cataloging practices.

WGBH and Pop Up Archive understand the need for large-scale digital indexing and analysis of audio to enhance its descriptive data and improve discoverability. Speech to text and audio analysis tools have been adapted for use with audio collections containing speech, but they do not generate perfect transcripts, and the qualitative data is not always clean enough for consistent indexing. Training such tools and refining their output requires human oversight and quality control. Harnessing the interest of the public and spreading the work over many volunteers may be a solution to the lack of resources available to describe this content. The challenge is to make the games for capturing the data easy, intuitive and engaging – tools that anyone could use or implement.

WGBH and Pop Up Archive are seeking $899,910 for a Research Grant to test such methods through the American Archive of Public Broadcasting (AAPB). AAPB, a collaboration between WGBH and the Library of Congress, is coordinating a national effort to identify, preserve, and make accessible a digital archive of public television and radio dating back to the late 1940s. Our research questions are: How can crowdsourced improvements to machine-generated transcripts and tags increase the quality of descriptive metadata and enhance search engine discoverability for audiovisual content? How can a range of web-based games create new points of access and engage the public engagement with time-based media through crowdsource tools? What qualitative attributes of audiovisual public media content (such as speaker identities, emotion, and tone) can be successfully identified with spectral analysis tools, and how can feeding crowdsourced improvements back into audio analysis tools improve their future output and create training data that can be publicly disseminated to help describe other audiovisual collections at scale?

The project will fund 1) **Audio Analysis tools** - development and use of speech-to-text and audio analysis tools to create transcripts and qualitative waveform analysis for almost 40,000 hours of AAPB digital files; 2) **Metadata Games** - development of open source web-based tools to improve transcripts and descriptive data by engaging the public in a crowdsourced, participatory cataloging project; 3) **Evaluating Access –** a measurement of improved access to media files from crowdsourced data; 4) **Sharing Tools** - open source code release for tools developed over the course of the grant, and 5) **Teaching Data Set**- the publication of initial and improved data sets to 'teach' tools and provide a public database of audiovisual metadata for use by other projects working to create access to audiovisual material.

The 2014 National Digital Stewardship Agenda recommends: "Engage and encourage relationships between private/commercial and heritage organizations to collaborate on the development of standards and workflows that will ensure long-term access to our recorded and moving image heritage."[1] These partnerships are critical in order to move the needle of audiovisual access issues of national significance. WGBH and Pop Up Archive are eager to continue building such a relationship so that the innovations in technology, workflows, and data analysis advanced by the private sector are fully and sustainably leveraged for U.S. public media and cultural heritage organizations.

1http://www.digitalpreservation.gov/ndsa/documents/2015NationalAgenda.pdf

IMLS Research Grant
Improving Access to Time-Based Media through Crowdsourcing and Machine Learning
WGBH Educational Foundation: American Archive of Public Broadcasting and Pop Up Archive
June 1, 2015

1. **Statement of Need**

This project will address the challenges faced by many libraries and archives trying to provide better access to their media collections through online discoverability. The project will combine technological and social approaches for metadata creation by leveraging scalable computation and engaging the public — the end users — to improve access through crowdsourcing games and tools for time-based media.

The problem of scale for audiovisual collections

Libraries and public media archives across the country house tens of millions of audio and video recordings,[1] of which spoken word recordings are a significant component. Given this scale, it is difficult for this content to be made accessible by standard cataloging practices. WGBH has faced this challenge directly in connection with its work on the American Archive of Public Broadcasting (AAPB). The AAPB, a collaboration between WGBH and the Library of Congress (the Library), is coordinating a national effort to identify, preserve, and make accessible as much as possible a digital archive of public television and radio dating back to the late 1940s. The initial collection consists of about 40,000 hours, or nearly 70,000 files, of digital media selected by more than 100 public media stations and organizations with little consistent descriptive data. The resources necessary to catalog large collections like AAPB's by standard cataloging practices are overwhelming and unattainable. For time-based media, the content needs to be seen or heard in real time for a human to be able to describe it. We have calculated that to 'lightly' catalog the AAPB collection of 70,000 digital items, only spending 15-20 minutes per item, it would take one person eight years. To fully catalog the collection, complete with subject headings, and name authorities, one should spend approximately one hour per program; it would take one person 32 years working full time.

Audiovisual media must be digital or digitized to be broadly accessible. However, just because an audiovisual collection is digital does not guarantee that it will be discoverable to scholars or the general public. A 2010 survey reported that 84% of searching on the web begins with search engines.[2] No matter how many hours of digital content are available on the web, without robust descriptive metadata, it is not discoverable. To improve discoverability and access for scholars and researchers, the content needs better descriptive information that can be indexed by search engines.

In addition, the rate at which born digital media is created increases every day. Twelve hours of audio are uploaded to SoundCloud every minute, and 300 hours of video are uploaded to YouTube every minute. Managing audiovisual content with attached descriptive metadata at this scale is a critical issue for the future of

---

1 In 2007, the Association of Research Libraries estimated that its 123 member libraries held more than 10 million audio recordings (National Recording Preservation Board,*The State of Recorded Sound Preservation in the United States: A National Legacy at Risk in the Digital Age*, http://www.clir.org/pubs/reports/pub148/pub148.pdf, p.10). A Heritage Health Index published in 2005 revealed that 46.4 million collection items of recorded sound were housed within the institutions that participated in a baseline cultural heritage study. See Heritage Preservation, Inc., *A Public Trust at Risk: The Heritage Health Index Report on the State of America's Collections*, 162.

2 OCLC. *Perceptions of Libraries, 2010*. http://www.oclc.org/content/dam/oclc/reports/2010perceptions/2010perceptions_all.pdf , p. 32.

collecting and access. If cultural heritage institutions are to keep up with the pace of audiovisual content creation, they need practices that can radically scale to meet the pace of creation.

Potential solutions using audio analysis tools
Creating transcripts of the audio is a potential solution to describe the content and expose the text to search engines. Speech to text, or audio to text, tools can be adapted or "trained" for use with specific materials to achieve greater degrees of accuracy than ever before, but they do not generate perfect transcripts, and the data is not always clean enough for consistent indexing. There is no easy method to correct inaccuracies for large quantities of media transcripts. Similarly, audio analysis tools have progressed far enough (and have been implemented for collections ranging from bird calls to poetry) that they could be trained and applied at scale to oral history and archival audio collections containing speech. However, training such tools and refining their output requires human oversight and quality control. Harnessing the interest of the public and spreading the work over many volunteers may be a solution to the lack of resources available to describe this content. The challenge is to make the tools for capturing the data easy, intuitive and engaging – tools that anyone could use or implement.

Computation + crowdsourcing to describe audiovisual content
WGBH and Pop Up Archive are seeking $899,910 for a Research Grant to test such methods through the American Archive of Public Broadcasting (AAPB). Through an iterative, design-based research approach we will explore the following research questions:

- How can crowdsourced improvements to machine-generated transcripts and tags increase the quality of descriptive metadata and enhance search engine discoverability for audiovisual content? How can a range of web-based games create new points of access and engage the public engagement with time-based media through crowd source tools? What qualitative attributes of audiovisual public media content (such as speaker identities, emotion, and tone) can be successfully identified with spectral analysis tools, and how can feeding crowdsourced improvements back into audio analysis tools improve their future output and create publicly available training data that can help with cataloging other audiovisual collections at scale?

This project will use content from the AAPB as the sample data set to answer our questions. The project will fund 1) **Audio analysis tools**: the development and use of speech-to-text and audio analysis tools to create transcripts and qualitative waveform analysis for approximately 40,000 hours of AAPB digital files; 2) **Time-based media crowdsourcing games**: the development of open source web-based games to improve transcripts and descriptive data by engaging the public in a crowdsourced, participatory cataloging project; 3) **Evaluating access:** a measurement of improved access to media files from crowdsourced data; 4) **Sharing tools** - open source code release for tools developed over the course of the grant, and 5) **Teaching data set**: the publication of initial and improved data sets to 'teach' tools and provide a public database of audiovisual metadata (National Archival Audio Fingerprint) for use by other projects working to create access to audiovisual material.

Building on current computational techniques
While computational approaches for cataloging and analyzing text and image-based collections are increasingly common, audiovisual collections have received less attention to date. WGBH and Pop Up Archive plan to build and improve on the work of COMMA[3], a BBC R&D project, and High Performance Sound Technologies for

3 http://www.bbc.co.uk/rd/projects/comma

Access and Scholarship (HiPSTAS), both of which have been developing new technologies for facilitating automated metadata extraction from time-based media collections. This emerging technology increasingly enables metadata to be created automatically for digital time-based media through programmatic analysis and speech-to-text software. Specifically, we are interested in building upon the BBC's and HiPSTAS's work to create extensive training data for speaker identification and other qualitative audio attributes, to compare automatically-extracted metadata to pre-existing taxonomies (e.g. DBPedia or other controlled vocabularies), and to provide a baseline workflow leveraging multiple audio analysis tools, crowdsourcing, and training data that can be used by other audiovisual collections.

Automatically generated metadata from speech can provide a map of content and topics addressed in US public broadcast material from the past 50 years and an authoritative index of nationally significant speaker voices. We seek to augment approximately 40,000 hours of U.S. public broadcast content with automatically generated metadata improved through crowdsourced contributions. This will serve as a national "jumping off point" for digital discovery and analyses of US public media and archival spoken word audio. Tools and methodologies will be shared with the academic and open source communities with the explicit goal of creating reusable workflows for similar endeavors.

Background on crowdsourcing techniques

There are a number of existing projects aimed at gathering metadata for large troves of digital resources. Examples include the Trove project, conducted by the National Library of Australia (http://trove.nla.gov.au/), Family Search (familysearch.org), the BBC World Service Archive (http://worldservice.prototyping.bbc.co.uk/), and the British Library Georeferencer (http://www.bl.uk/maps/). When it comes to time-based media, a number of fields of data can be gathered by end users without their having to watch or listen to the full length of a program. For the crowdsourced component, we will build on and contribute to existing open source Metadata Games tools developed by TiltFactor at Dartmouth College. We will launch a platform that guides people through a series of steps to enable metadata tagging without requiring existing knowledge of metadata tagging. This project will test the viability of using the public and simple metadata gathering tools to add descriptive data to records of digitized media to increase access and discoverability.

The crowdsourcing project that most closely matches AAPB's needs is Dartmouth's Metadata Games effort (metadatagames.org). Code for this initiative has been shared with the open source community and can serve as an excellent foundation for our work. However, current games that work well with image-based collections do not entirely align with our needs, which are twofold: (1) we need human-generated and verified "tags" for audio segments that we can use to train classification and clustering algorithms for audio analysis, and (2) we need to verify and correct machine-generated transcripts by comparing that text to a specific portion of a media file. In addition, to date, the data generated from Metadata Games does not get entered into the authoritative metadata repository. In order to automatically funnel user-generated metadata into the AAPB metadata repository, we will build upon relevant code developed by Metadata Games, tap their expertise with games and users, and share our revised code with them and the open source community to further refine their code.

Urgency of need

As leaders in audiovisual preservation, access, and analysis of culturally significant audiovisual material, WGBH and Pop Up Archive are continually confronted with the need for large-scale digital indexing and analysis of audio to enhance its descriptive data and improve discoverability.

Pop Up Archive has fostered partnerships with dozens of partners at libraries, archives, and public media organizations to index almost 1,000,000 minutes of recorded sound since 2012 including over 10,000 audio items preserved at the Internet Archive (archive.org). Pop Up Archive was created in response to these organizations' desire to create access to their audiovisual collections by facilitating cataloging, search, and public engagement at scale, in spite of a general lack of knowledge and technical capability for handling audiovisual content. Pop Up Archive's partners' collections include 7500 hours of archival Studs Terkel broadcasts from the WFMT Radio Network, tens of thousands of oral histories from across the United States collected by StoryCorps, the New York Public Library, and numerous university oral history initiatives. Funded by the John S. and James L. Knight Foundation and National Endowment for the Humanities, Pop Up Archive works in collaboration with the Public Radio Exchange and under the advice of the British Broadcasting Corporation R&D division. Broadcasting since 1951, WGBH currently maintains an archive of over 750,000 items and the public website OpenVault. In collaboration with the Library, WGBH is a permanent steward of the AAPB initiative and collection, which seeks to preserve and make accessible a digital archive of publicly funded television and radio broadcasting created by stations and organizations in communities across America.

Together, WGBH and Pop Up Archive represent the pinnacle of public media service and the vanguard in digital audio innovation. The 2015 National Agenda for Digital Stewardship Agenda sets forth as an actionable recommendation the following: "Engage and encourage relationships between private/commercial and heritage organizations to collaborate on the development of standards and workflows that will ensure long-term access to our recorded and moving image heritage."[4] These partnerships are critical in order to move the needle on audiovisual access issues of national significance. WGBH and Pop Up Archive are eager to continue building such a relationship so that the innovations in technology, workflows, and data analysis advanced by the private sector are fully and sustainably leveraged for U.S. public media and cultural heritage organizations.

2. Impact

This project will develop workflows that employ emerging open source technology to create large quantities of metadata for digitized audiovisual content through highly accurate speech-to-text software, semantic tagging, and digital waveform data. It will result in a database of time-stamped transcripts and keyword tags for the entirety of the AAPB's approximate 70,000 digital files, as well as open source crowdsourcing tools for the public to enhance automatically-generated audiovisual metadata. It will also create the National Archival Audio Fingerprint: a public repository of training data consisting of samples of various audio attributes, including nationally significant audio "voice signatures," and an open source codebase of tools to enable other archives to conduct analyses of their audio collections. Through time-based media metadata games and tools, crowdsourcing participants will verify and correct transcripts and audio traits such as speaker identities generated from audio analysis tools to create training data for reuse as part of the audio fingerprint.

The tools and workflows we develop to answer our research questions will contribute to the National Digital Platform by addressing fundamental concerns for libraries and archives trying to scale cataloging and creation of descriptive metadata for audiovisual collections. For archives with large audiovisual collections, the project will demonstrate how speech-to-text tools can be used at scale to create transcripts and how waveform analysis software can be used to create metadata about other audio attributes. Our research will explore how this metadata can be augmented to improve access and systematized through workflows to more efficiently make large audiovisual collections accessible and discoverable. Pop Up Archive has already identified and

4 http://www.digitalpreservation.gov/ndsa/documents/2015NationalAgenda.pdf, p. 20.

contributed to the development of state-of-the-art speech-to-text technology from institutions such as the International Computer Science Institute in Berkeley, CA, in addition to scoping the requirements for augmenting the HiPSTAS platform to support public media content. As a result of this project, Pop Up Archive will build on these fundamental technologies by creating the audio fingerprint, a taxonomy for public media and culturally significant archival audio, and a comparative analysis of AAPB's 40,000 hours of recorded sound.

This project will aim to produce open source tools that can shortcut the audiovisual cataloging process and provide at least minimal data for meaningful searchability. Not only will potentially tens of thousands of hours of audiovisual content become more accessible to the public, but tools, training data, and workflows will be created that can be improved upon and widely implemented. Our research will identify low-cost solutions for archives and libraries to provide better access to their audiovisual collections and expose more data to search engines. The project will improve open source audio analysis tools (see Project Design) by providing vocabularies, models, and training data unique to archival content that have been amassed through crowdsourced contributions designed to "teach" the tools through verified training data.

Several iterative research phases are built into the project as part of our design-based research approach. Our questions address how the methodology can best improve access to audiovisual collections and enable other libraries and archives to benefit from the same technology. Throughout the development of the time-based media crowdsourcing tools at WGBH, the project team will conduct multiple usability tests to ensure intuitive usability of the tools. We will measure improvement in access by conducting an evaluation to see if researchers using the AAPB digital archive website can find what they want with minimal metadata. We will conduct a summative evaluation at the end of the project to test whether machine-generated transcripts and metadata with crowdsourced improvements have improved the discoverability of the same objects. We will measure increased access to the content online as evidenced through user engagement, including monitoring web analytics for how much organic traffic is driven to the collection by direct transcript and metadata hits, in particular for items that have been cleaned up or augmented through crowdsourcing. We will ensure that the code we build on top of TiltFactor and HiPSTAS' existing code can be integrated into reusable workflows by working closely with those teams to identify and communicate with libraries and archives that would benefit from such workflows.

Near the end of the project, the project team at WGBH will also evaluate whether participants in crowdsourcing for time-based media enjoy the experience. Would they return to contribute again? What were the challenges they faced with while using time-based media versus other types of crowdsourcing games? Code and documentation usability and implementation will be measured by how often the code is downloaded, re-used, and modified once it is posted on Github. Our goal will be to illustrate that highly accurate transcripts and user-generated descriptive metadata increase access. The project will provide data and information to further improve tools for speech-to-text, audio analysis, and crowdsourcing games, including speech-to-text and waveform analysis training data.

It is paramount to the success of this project that the tools and workflows we create be usable by other audiovisual collections of national cultural significance, including those with few resources. Our collaborative relationships with TiltFactor, HiPSTAS, Pop Up Archive and the hundreds of organizations that partner with them, are critical steps toward accomplishing this goal. The project will raise the profile of audiovisual collections in the eyes of the general public because the materials will be more accessible through search engines like Google. Also, because the public will have played a part in making the content more easily discoverable, they will have become more aware of the potential treasures in archival audiovisual collections.

WGBH and the Library are committed to sustaining the AAPB collection and making as much of it available to the pubic as possible. As we continue to grow the collection, we will be able to reuse the tools to enhance future content through transcripts and user-generated metadata. The tools we create will be available on Github as open source repositories that can be built upon by the Pop Up Archive, TiltFactor, and HiPSTAS teams, their partners, and other libraries and archives through an MIT or similar open license.

This project will test the viability of using scalable computational means, the public crowdsourcing efforts, and simple metadata gathering tools to add descriptive data to records of digitized video and audio to increase access and discoverability. We will be informed by the experiences of other audio collections that have successfully employed HiPSTAS' software, as well as games through which TiltFactor has successfully facilitated crowdsourcing to gather metadata. We plan to have several face-to-face meetings with the TiltFactor and HiPSTAS teams as well as conduct evaluations throughout the project via usability studies, focus groups and surveys with potential users/game players, as well as with the researchers whose work will be affected by crowdsourced metadata.

Training data (the audio fingerprint) and tools will be published under a Creative Commons or similar license as a stand-alone product in multiple locations, such as Github. We will also work with an advisory board composed of leaders in national archival, library, broadcast, and digital stewardship efforts, including Emily Gore (Digital Public Library of America), and Roger MacDonald (Internet Archive), to widely disseminate the products of this grant.

3. Research Design

In order to answer our key research questions, this Project will be implemented in five phases. We will follow principles of design-based research to build and test each phase and product as we move through the project.

Phase One | Pre-Development Research (Months 1-2): WGBH will hold a virtual focus group among organizations conducting crowdsourcing initiatives to gain insight into the planning, technological considerations and outreach conducted by previous initiatives. WGBH will evaluate existing crowdsourcing tools, including those involving transcripts, as well as those with audiovisual components. In conducting this evaluation, WGBH will identify the functional requirements of crowdsourcing transcript improvement and tagging for audio and video collections. WGBH will conduct a pre-project evaluation of discoverability of content in the AAPB collection, which will be used to measure improvement of access with crowdsourced data at the end of the project.

Phase Two | Generate transcripts and conduct Adaptive Recognition with Layered Optimization (ARLO) analysis (Months 3-15): WGBH will prepare and deliver video and audio files to Pop Up Archive. Pop Up Archive has made great strides in automatic speech technology by working with language models trained for specific types of historic archival content. Pop Up Archive will process the AAPB digital files to generate time stamped transcript text and semantically analyze the text to create a database of entities. These transcripts will be made available in standard formats (txt, json, xml, srt)) through WGBH to scholars and public media institutions. To further accomplish the goal of bringing cutting-edge speech recognition methods to the GLAM (Galleries, Libraries, Archives, and Museums) community, Pop Up Archive will release standalone open source Kaldi speech-to-text software trained specifically for public media and implementable by any organization with the resources to download and install the software and support the processing of their own audio files. The Pop Up Archive platform (web service) will continue to support smaller organizations and organizations that are not compelled to install open source speech recognition technology on their own.

Pop Up Archive will work closely with Tanya Clement and the project team at HiPSTAS to implement the ARLO software suite on a to-be-determined subset of AAPB digital files. The goal of this step is to generate additional programmatic data about the audiovisual material beyond the actual words being spoken in the recordings. This work will first entail the identification of audio "qualities" other than speech, such as speaker identities, applause, laughter, and more nuanced vocal tones. This training data, or "audio fingerprint," will be publicly released for use by others in the GLAM community. Once this training data has been created, we will leverage ARLO to analyze a subset of AAPB digital files and programmatically generate additional qualitative metadata about that content. This analysis will be an iterative process as we share initial results with the scholarly and computer science communities and solicit their feedback and suggestions on which data is most useful to them.

Phase Three: Develop Crowdsourcing Transcript Tool (Months 5-17): Under the advice of Mary Flanagan and TiltFactor, WGBH and the HiPSTAS team will develop open source web-based tools to improve transcripts and descriptive data by engaging the public in a crowdsourced, participatory cataloging project. First, WGBH and the HiPSTAS team will determine how the tools will interact with existing infrastructure, including the AAPB metadata repository ("the Repository") and the public facing website at americanarchive.org ("the Website"). WGBH Digital and the HiPSTAS team will design templates for displaying the transcripts on the Website alongside spectograms generated by ARLO. MLA Developers will establish a path for ingestion of transcripts generated by Pop Up Archive into the Repository and Website. Once the path is developed, Pop Up Archive will begin sending transcripts to WGBH on a monthly basis, and MLA Developers will ingest the original transcripts into the Repository and Website. MLA developers will use the ARLO API to send human-generated audio "tags" for machine learning analyses, the results of which will be systematically validated by the HiPSTAS, WGBH, and Pop Up Archive teams before being ingested into the Repository and Website.

Building on the work of Metadata Games, WGBH Digital will begin development of the transcript tools by creating prototype wireframes. MLA Project Manager will conduct an explorative usability test with diverse users on the visual layout and interface element design of the tools, their logical flow, and the preferred behavior of users. Based on this feedback, WGBH Digital will modify wireframes and develop the transcript tools. MLA Project Manager will conduct an assessment usability test midway in the tool's development phase to evaluate the technology. This test will involve real-time trials of the tool to determine satisfaction, effectiveness, and overall usability. Based on feedback from a diverse user group, WGBH Digital will continue development and improvement of the transcript tools. Throughout the course of development, the MLA Project Manager will monitor development of the tools and report bugs and issues.

Phase Four | Launch of Games (Months 17-26): MLA Project Manager will create instructional guides and a video for users, which will accompany the transcript tools on the web. WGBH will open the platform to an initial user base, requesting feedback on the tools and instructional documentation. MLA Project Manager will conduct outreach to engage users across the country. MLA will disseminate the tools by sharing a press release, posting information about the tools on listservs including the SAA, AMIA, NEA, and H-Net (a listserv for humanities professionals and scholars), and sharing the tools with scholars including those involved with the Library of Congress' Radio Preservation Task Force. MLA Project Manager will collaborate with WGBH's National Productions (which includes award-winning series such as *Antiques Roadshow*, *Frontline*, *NOVA*, and *American Experience*), to educate them on ways to promote and share crowdsourcing tools through their broad social media networks. The Project Manager will give a presentation at WGBH's Social Media Meet-Up, attended by local and national social media and digital producers, to inform them about the tools and ways to

incorporate it into their digital and social media strategies. More than 100 AAPB participating organizations will be able to share the tools with their communities. Throughout the project, the Project Team will present developments and highlights at professional conferences, such as Society of American Archivists, Association of Moving Image Archivists, and the PBS Annual Meeting. The Project Team will also work with TiltFactor and HiPSTAS to run workshops at these events to teach other collections how to use the tools.

Phase Five | Evaluation and Dissemination of Findings (Months 25-30): At the beginning of this phase, MLA developers will evaluate the data contributed by users through the process of indexing it into Solr. Developers will share information with the MLA Project Manager about the consistency and value of the crowdsourced data. Once ingested into Solr and made available and searchable on the Website, MLA Project Manager will conduct a summative evaluation to measure improved access to media files with the added crowdsourced data. WGBH will create a survey that will be sent out to listservs such as Code4Lib, requesting other library/archives developers to evaluate the source code documentation and installation process. MLA Project Manager will create a survey to distribute to users to evaluate the crowdsourcing experience overall. Using the results of the evaluation, WGBH will write a final report for the project, which will be shared with the library and archives community. WGBH and Pop Up Archive will release any other open source software and datasets that have not yet been released.

WGBH Developers in Digital and MLA will write and disseminate documentation to accompany the open source code for the transcript tools. WGBH will share open source code developed through the project with the community by releasing it on Github and pursuant to an MIT license or similar open license.

In addition to open source software tools leveraged and improved upon as part of the IMLS National Leadership Research Grant (ARLO, Metadata Games, Kaldi), WGBH and Pop Up Archive will release the following data outputs resulting from the project:

- Time-stamped transcripts and semantic entities derived through speech recognition
- Training data for the ARLO audio analysis tools and ultimate resulting analysis of AAPB content
- Improvements to transcripts and semantic entities generated through crowdsourcing
- Additional qualitative metadata created through crowdsourced human cataloging efforts

5. Project Resources: Personnel, Time, Budget: Key Project Staff:

Karen Cariani (10% time for 30 months), Director of the WGBH Media Library and Archives, will lead the project. Cariani has been director of the Media Library since its inception in 1990. She co-led the AMIA Local Television Case Studies and Symposium Task Force from 2001-2004. She has been project director for WGBH Teachers' Domain initiative, WGBH Open Vault, WGBH Mellon Digital Library project, the American Archive Content Inventory Project, and the American Archive of Public Broadcasting, and directed the development and testing of the WGBH DAM system. She will supervise the administrative operations of the project to ensure it remains on time, on budget, and within the scope of the proposal.

Casey Davis (see budget for detail on time percentages over the course of the project) is currently the project manager for the American Archive of Public Broadcasting. She will oversee the project's day-to-day activities and collaborate with the WGBH Digital team, the Project Advisory Board, AAPB participating organizations, and potential users of the crowdsourcing tools.

Nick Pollard (5% time for 30 months), is a Coordinating Producer, overseeing all finance and operations for

the Media Library and Archive and grant funded projects. Prior to joining WGBH in 2012 he worked in various finance and operations roles, overseeing budgets and operations, as well as investment management for private and public companies. Nicholas has an MBA from the Boston University Graduate School of Management, and a BA from Hobart & William Smith Colleges. He will oversee the project's finances and budget.

Pop Up Archive, Anne Wootton, CEO, (25% time for 24 months): Anne will oversee all computational aspects of the grant, including development and implementation of speech-to-text software, audio waveform processing via ARLO, and packaging/dissemination of related outputs. Anne's responsibilities for the IMLS National Leadership Research Grant will dovetail with her full-time role as co-founder of Pop Up Archive, where her mission is to bring cutting-edge computational cataloging methods to U.S. audiovisual collections small and large. Anne is a recipient of the Knight News Challenge: Data, a member of the National Digital Stewardship Alliance Innovation Working Group, and former chair of the AMIA PBCore Subcommittee's Education Team. She was a judge of the 2014 Libraries cycle of the Knight News Challenge and has presented at conferences targeted at GLAM professionals on how to make audio discoverable by search engines as well as scalable computational methods for treating audiovisual media.

HiPSTAS Developer (25% time for 12 months) This developer will be responsible for hooking up ARLO to a supercomputer cloud server provided by Pop Up Archive as well as overseeing the ingestion of up to 40,000 hours of WGBH sound files. Responsibilities will also include Java and Python development for improving ARLO, web development for ARLO's coordination with Metadata Games, and the implementation of the ARLO data API.

UT iSchool Graduate students (10 hrs/week for 12 months) Graduate students will generate tags with WGBH sound files, validate machine learning results, update and manage user documentation, create training videos, and work with the MLA Project Manager and Pop Up Archive to help assess the Metadata Games interface, create tags, and clean up transcripts as well as gather user feedback on the crowdsourcing experience overall.

Advisors

Mary Flanagan (Dartmouth College, TiltFactor) TiltFactor is a game design studio at Dartmouth dedicated to crafting and studying games and playful solutions for social impact. TiltFactor developed Metadata Games, the National Standard open source crowdsourcing game platform. Mary will advise WGBH developers on the reuse of Metadata Games open source code, user interfaces, and user needs for crowdsourcing tools. She will collaborate with WGBH and Pop Up Archive to share output with relevant academic communities and solicit their feedback.

Tanya Clement (UT-Austin, HiPSTAS) The HiPSTAS project objective is to develop a virtual research environment in which users can better access and analyze spoken word collections of interest to humanists. Tanya will oversee subcontracted developers from UT-Austin and advise Pop Up Archive in the use of ARLO, including iterating on training data and output generated through the software. She will also collaborate with Pop Up Archive and WGBH to share output with relevant academic communities and solicit their feedback.

Roger MacDonald (Broadcast TV News Archive, Internet Archive) The Broadcast TV News Archive is a research library service intended to enhance the capabilities of journalists, scholars, teachers, librarians, civic organizations and other engaged citizens by repurposing closed captioning to enable users to search, quote and borrow U.S. TV news programs. Roger will share experience and lessons learned through the development of the Broadcast TV News Archive, including sharing best practices for making time-based audiovisual content

accessible and understandable to a wide range of end users such as journalists, scholars, and the general public, and offering feedback on strategies for engaging a broad community in crowdsourced metadata cleanup and description.

Emily Gore (Digital Public Library of America) The DPLA brings together the riches of America's libraries, archives, and museums, and makes them freely available to the world. Emily will share experience and lessons learned through the development of the DPLA, including challenges faced integrating audiovisual records and materials and navigating copyright concerns as they pertain to inconsistent rights statements and the abilities of audiovisual collections to effectively participate in national archival efforts.

6. Communication Plan

The project communication plan will target the following audiences: crowdsourcing tool users, AAPB digital archive users, and the professional archival community. The project's communication plan will benefit from the broad collaborations developed throughout the project: with Pop Up Archive; the project Advisory Board; crowdsourcing tool users; usability study participants, including scholars, novice users and beta test centers; public media stations; and within WGBH's Media Library and Archives and Digital departments.

Evaluation: Assessment of needs will occur at the beginning of the project through focus groups, evaluation of existing tools, and functional requirements analysis. The project team will evaluate the tools during development through usability studies of crowdsourcing tools, as well as conducting multiple analyses of waveform data and sharing that data with stakeholders in computer science, humanities research, audiovisual collections, and public media to generate and refine the national audio fingerprint.

In the final phase of Evaluation, the Project Team will evaluate the success of the project in a variety of methods: 1) evaluation the data contributed by users through the process of indexing it into Solr; 2) summative evaluation via web analytics and user survey to measure improved access to media files with the added crowdsourced data; 3) evaluation of crowdsourcing experience by crowdsourcing participants via surveying; and 4) evaluation of ease of installation and implementation through request of community feedback.

Outreach and Promotion: The Project Team will be highly involved in outreach and dissemination of the audio analysis, crowdsourcing tools, and documentation. The project will solicit participation from a variety of stakeholders to build a community of constructive interest. This may lead to increased use and awareness of the tools among potential users, potential researchers of the AAPB collection, and other institutions that might benefit from installing the tools locally. The tools will be shared with potential users via listservs and social media. The project progress and accomplishments will be publicized. A press release will be posted for major print publications, professional organization journals, and listservs. The AAPB team will propose panel presentations to share findings and results at professional library, archives, and educational conferences, such as AMIA, SAA, NEA, and PBS.

Technical Documentation: All reports and documentation will be posted on the AAPB website. A project blog will be written to highlight progress, as well as different approaches to engaging with the tools individually and at organizations seeking to implement the program into their coordinated activities. All code, documentation and instructional materials for installation and implementation of the crowdsourcing tools will be open source and shared on a variety of platforms, including Github.

Improving Access to Time-Based Media through Crowdsourcing & Machine Learning

IMLS National Leadership Grants for Libraries: Research Grant
WGBH Educational Foundation, American Archive of Public Broadcasting, and Pop Up Archive
Schedule of Completion

Green = WGBH Project Manager
Blue = WGBH MLA Developer
Yellow =WGBH Digital developer
Red = Pop Up Archive + HiPSTAS

Phase 1: Pre-Development Research	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**26**	**27**	**28**	**29**	**30**
WGBH holds focus groups with organizations conducting crowdourcing																														
WGBH evaluates existing crowdsourcing tools & discoverability of AAPB collection																														
WGBH identifies functional requirements of crowdourcing transcipts for video and audio																														
Phase 2: Generate transcripts and conduct Adaptive Recognition with Layered Optimization (ARLO) analysis																														
WGBH prepares and delivers video and audio files to Pop Up Archive																														
Prepare training data for ARLO qualititive analysis of audio waveforms																														
Process 38,000 hours of content to generate timestamped machine transcripts and metadata																														
Identify audio "qualities" such as speaker identities, applause, laughter, tone																														
Run initial ARLO analysis on subset of AAPB content for national audio fingerprint																														
Train language models for Kaldi speech-to-text software																														
Share ARLO analysis for feedback and iteration																														
Release ARLO training data																														
Release open-source speech-to-text software																														
Phase Three: Develop Crowdsourcing Transcript Tool																														
Determine how time-based tool will interact with AAPB website																														
Develop template for displaying time-based media, transcripts and metadata on AAPB website																														
Developing the path for ingest into AAPB metadata repository																														
As transcripts and metadata are delivered to WGBH by Pop Up Archive, WGBH ingests transcripts and metadata into AAPB website																														
Develop prototype wireframes for time-based media crowdsourcing tool																														
Conduct exploratorive usability test																														
Modify wireframes and develop tool																														
Conduct assessment usability test																														
Continue development and improvement of tool																														
Monitor development of tool and report bugs and issues																														
Phase Four: Launch of Games																														
Create instructional guides and video for users to accompany time-based media tools on the website																														
Open crowdsourcing platform to initial user base																														
Conduct outreach to engage users across the country																														
Re-run ARLO analysis and adaption of national audio fingerprint based on feedback																														
Write and disseminate documentation for time-based media crowdsourcing tool code and installation																														
Phase Five: Evaluation and Dissemination of Findings																														
Ingest data generated by users into SOLR index																														
Evaluate data gathered by crowdsourcing transcript corrections																														
Evaluate improved discoverabilty among users																														
Evaluate documentation for installation of code																														
Evaluate experience of volunteers																														
Write final report																														
Release any open-source software and datasets not yet released																														

OMB No. 3137-0071, Exp. Date: 09/30/2015

IMLS Research Grant
Improving Access to Time-Based Media through Crowdsourcing and Machine Learning
WGBH Educational Foundation: American Archive of Public Broadcasting and Pop Up Archive
DIGITAL STEWARDSHIP SUPPLEMENTARY INFORMATION FORM

Introduction:
IMLS is committed to expanding public access to IMLS-funded research, data and other digital products: the assets you create with IMLS funding require careful stewardship to protect and enhance their value. They should be freely and readily available for use and re-use by libraries, archives, museums and the public. Applying these principles to the development of digital products is not straightforward; because technology is dynamic and because we do not want to inhibit innovation, IMLS does not want to prescribe set standards and best practices that would certainly become quickly outdated. Instead, IMLS defines the outcomes your projects should achieve in a series of questions; your answers are used by IMLS staff and by expert peer reviewers to evaluate your proposal; and they will play a critical role in determining whether your grant will be funded. Together, your answers will comprise the basis for a work plan for your project, as they will address all the major components of the development process.

Instructions:
If you propose to create any type of digital product as part of your proposal, you must complete this form. IMLS defines digital products very broadly. If you are developing anything through the use of information technology – e.g., digital collections, web resources, metadata, software, data– you should assume that you need to complete this form.

Please indicate which of the following digital products you will create or collect during your project. Check all that apply:

	Every proposal creating a digital product should complete …	Part I
	If your project will create or collect …	**Then you should complete …**
X	Digital content	Part II
X	New software tools or applications	Part III
X	A digital research dataset	Part IV

PART I.

A. Copyright and Intellectual Property Rights

We expect applicants to make federally funded work products widely available and usable through strategies such as publishing in open-access journals, depositing works in institutional or discipline-based repositories, and using non-restrictive licenses such as a Creative Commons license.

A.1 What will be the copyright or intellectual property status of the content you intend to create? Will you assign a Creative Commons license to the content? If so, which license will it be? http://us.creativecommons.org/
All code will be released open-source under an MIT License or similar license.
All datasets will be released under the (CC BY-NC 4.0) Attribution-NonCommercial 4.0 International license.

A.2 What ownership rights will your organization assert over the new digital content, and what conditions will you impose on access and use? Explain any terms of access and conditions of use, why they are justifiable, and how you will notify potential users of the digital resources.
WGBH and Pop Up Archive will each own components of the digital content (code and research data) created through this grant. Code developed for speech-to-text, audio waveform analysis, and crowdsourced metadata games will all developed from previous open-source efforts and final products will also be open-sourced.
We will restrict use of the code and research data to non-commercial applications. The Project Team will be highly involved in outreach and dissemination of the audio analysis and crowdsourcing tools and accompanying documentation to potential users of the tools including researchers other institutions that might benefit from installing the tools locally to engage in computational and crowdsourcing efforts to improve access to their audiovisual collections. The tools will be shared with potential users via listservs and social media, and a press release to major print publications, professional organization journals, and listservs. The Project Team will also propose panel presentations to share findings and results at professional library, archives, and educational conferences, such as AMIA, SAA, NEA, and PBS.

A.3 Will you create any content or products which may involve privacy concerns, require obtaining permissions or rights, or raise any cultural sensitivities? If so, please describe the issues and how you plan to address them.
No.

Part II: Projects Creating Digital Content

A. Creating New Digital Content

A.1 Describe the digital content you will create and the quantities of each type and format you will use.
We are not creating digital content for this project other than the digital tools and research data described in detail in Parts III and IV. However, we are working with audio digitized as part of the American Archive of Public Broadcasting initiative, initially funded by the Corporation for Public Broadcasting and overseen today by WGBH and the Library of Congress. The digital content that comprises the American Archive of Public Broadcasting consists of audiovisual content collected from public radio and television stations across the United States.

A.2 List the equipment and software that you will use to create the content or the name of the service provider who will perform the work.

The transcripts will be created using the Kaldi speech-to-text software and the output will be available as timestamped json, xml, txt, and srt.

A.3 List all the digital file formats (e.g., XML, TIFF, MPEG) you plan to create, along with the relevant information on the appropriate quality standards (e.g., resolution, sampling rate, pixel dimensions).

json, xml, txt, and srt

B. Digital Workflow and Asset Maintenance/Preservation

B.1 Describe your quality control plan (i.e., how you will monitor and evaluate your workflow and products).

For transcription and tag generation, Pop Up Archive imports files in CSV batches onto an Amazon Web Services server where they are queued for processing by the Pop Up Archive system. Pop Up Archive monitors all uploads, transcribing, transcription, and automated metadata creation via a preexisting internal dashboard. Email alerts can be set to dispatch upon completion of processing for each file. Failed attempts are automatically requeued and retried. Pop Up Archive uses best-of-breed industry tools like New Relic and Logentries to constantly evaluate and improve transcript creation workflow, identify problems and optimize turnaround times.

B.2 Describe your plan for preserving and maintaining digital assets during and after the grant period (e.g., storage systems, shared repositories, technical documentation, migration planning, commitment of organizational funding for these purposes). Please note: Storage and publication after the end of the grant period may be an allowable cost.

All transcripts will be stored locally on LTO-6 tape in their original format (XML, JSON, and TXT). Additionally, all transcripts will be added to the American Archive of Public Broadcasting authoritative metadata repository, the Archival Management System (AMS). Hosted on a dedicated server at OVH in Canada, the AMS provides search and discovery access to all AAPB metadata, including descriptive, technical and preservation metadata. The metadata is stored in a mysql database based on the PBCore and PREMIS data models, which is backed up daily to a second server hosted by Amazon. In addition to preserving transcripts in the AMS and locally, all transcripts will be made discoverable by a Solr index and exposed on the public-facing AAPB blacklight application at americanarchive.org.

C. Metadata

C.1 Describe how you will produce metadata (e.g., technical, descriptive, administrative, preservation). Specify which standards you will use for the metadata structure (e.g., MARC, Dublin Core, Encoded Archival Description, PBCore, PREMIS) and metadata content (e.g., thesauri).

The metadata that will be produced through this project will consist of transcripts for all digitized video and audio in the AAPB collection and user-generated tags.

Transcripts will be produced and delivered to WGBH in json, txt, srt, and xml. The xml schema used is the W3C Transcript standard. As volunteers provide crowdsourcing contributions to fix the transcripts via the crowdsourcing tool developed by WGBH, all updated transcripts will remain in the same format.

As crowdsourcers fix transcripts, they will be able to generate descriptive tags about the content they are viewing or listening to. All tags will be delivered via an API and will be added to the AAPB authoritative metadata repository, the AMS, which is built on the PBCore data model, within the pbcoreSubject field. AAPB staff will use a pbcoreSubject @type attribute to indicate that the tag was user-generated. When exposed on the public-facing AAPB website, user-generated tags will be displayed alongside staff-generated descriptive metadata, but the user-generated pbcoreSubjects will indicate that the metadata was user-generated. The PBCore format will provide the most precise format for metadata interchange, and will allow this metadata to be converted into other standard formats (METS, Dublin Core, etc) as needed.

C.2 Explain your strategy for preserving and maintaining metadata created and/or collected during your project and after the grant period.
The user-generated tags will be stored along with other official metadata in the AMS. The transcripts will be preserved by storing copies in two locations. One copy will be stored on a local WGBH media

server, and another copy will be stored on Amazon S3. The metadata collected during the project will be added to the American Archive catalog and exposed to the public via the American Archive website. As part of the AAPB collection it will be preserved at the Library of Congress along with other metadata in their collection as per the mission and mandate of the Library.

C.3 Explain what metadata sharing and/or other strategies you will use to facilitate widespread discovery and use of the digital content created during your project (e.g., an Advanced Programming Interface, contributions to the DPLA or other support to allow batch queries and retrieval of metadata).

For user-generated tagging, we will use PBCore, which is based on Dublin Core and can be used in conjunction with METS/PREMIS and mapped to other descriptive standards (MODS, Dublin Core, EBUCore, etc.) for data exchange, re-use, and discoverability. The transcripts and PBCore data will be exportable and harvestable through an OAI feed, allowing participation in the Digital Public Library of America and other portals of digital libraries.

D. Access and Use

D.1 Describe how you will make the digital content available to the public. Include details such as the delivery strategy (e.g., openly available online, available to specified audiences) and underlying hardware/software platforms and infrastructure (e.g., specific digital repository software or leased services, accessibility via standard web browsers, requirements for special software tools in order to use the content).

Depending on the particular content, the AAPB will be able to provide the following types of access and use of transcripts according to rights already obtained from the contributing organizations, Section 108 (rights of libraries and archives) and Section 107 (fair use) of Title 17 U.S.C.:

a) Provide on-location access to all transcripts at WGBH and the Library of Congress,

b) Provide access to all transcripts for material that is in the AAPB Online Reading Room,

c) Index all transcripts into Solr to allow for improved searching on the AAPB website for all content in the collection, and/or

d) Allow for crowdsourcing participants to fix and improve all transcripts created through this project by having crowdsourcers fix segments of transcripts when watching shortened clips of its corresponding program.

The AAPB website is located at americanarchive.org. The site is a Ruby on Rails application, using Blacklight to access a Solr index, and running on AWS EC2. The records in Solr are created from PBCore XML files that are periodically harvested from an HTTP API provided by our authoritative

metadata repository, the Archival Management System (AMS). The AMS is an open-source application developed by Audiovisual Preservation Solutions and is built on the PBCore and PREMIS data models. The AMS is a php application and uses a mysql database to store tables of metadata. Source code for the AMS is available at https://github.com/avpreserve/ams. Source code for the AAPB website is available at https://github.com/wgbh/aapb2. The AAPB website is accessible and useable in a variety of operating environments and browsers, including Google Chrome, Mozilla Firefox, Safari and Internet Explorer. The AAPB website design is responsive and is accessible via mobile device, tablet and desktop/laptop.

D.2 Provide URL(s) for any examples of previous digital collections or content your organization has created.

americanarchive.org

Part III. Projects Creating New Software Tools or Applications

A. General Information

A.1 Describe the software tool or electronic system you intend to create, including a summary of the major functions it will perform and the intended primary audience(s) the system or tool will serve. We intend to create two primary software tools intended for national use by institutions and/or collections of digital audiovisual content:

1) Tools for computational analysis and metadata creation for digital audiovisual content
    - The tools will generate timestamped machine transcripts and descriptive tags
    - The tools will be capable of analyzing waveforms for qualitative attributes such as speaker identities and non-speech characteristics such as applause or laughter.
2) Tools to improve transcripts and descriptive data by engaging the public in a crowd-sourced, participatory cataloging project

    - A metadata game that uses game mechanics to encourage users to add meta-tags to transcripts and descriptive data.
    - Further extensions of the games found at http://www.metadatagames.org/, customized for the specific needs of the AAPB.

A.2 List other existing digital tools that wholly or partially perform the same functions, and explain how the tool or system you will create is different.

1) Speech-to-text software such as the open-source libraries from CMU Sphinx and Kaldi exists, but there is no open-source, publicly available speech-to-text software trained specifically for

use with archival, oral history, or broadcast audio. Similarly, waveform analysis software has focused largely on music, and the ARLO software developed by HiPSTAS (which this project will build upon) has focused to date on recorded sound such as bird calls or poetry.

2) While great strides with crowdsourced metadata games have been made in recent years through projects such as Tilt Factor's Metadata Games (which this project will build upon), the games have focused primarily on text and images and have not been developed for time-based media.

B. Technical Information

B.1 List the programming languages, platforms, software, or other applications you will use to create your new digital content.

The transcription software will be build on Kaldi. The semantic analysis and tagging components will leverage a combination of OpenCalais, the Yahoo Analysis API, DBPedia, and Mango (pending its open-source release by the BBC). The waveform analysis software will be a specially-trained instance of HiPSTAS' ARLO software.

WGBH will use a combination of Python and Unity to create the time-based crowdsourcing tool. Any back end tool would be created in Django. We also intend to make use of the Metadata Games open source code as a basis for our work.

B.2 Describe how the intended software or system will extend or interoperate with other existing software applications or systems.

All of the software will be available open-source and standalone, so it could be downloaded and installed by any institution with resources and incentive to do so. In addition, Pop Up Archive and HIPSTAs provide web-based services or interfaces that enable access to transcription and waveform analysis software for institutions less inclined to install and run the code themselves.

WGBH Digital will create the time-based media crowdsourcing tool and will provide WGBH MLA with a REST API to query and gather updated transcripts in the W3C Transcript XML format.

B.3 Describe any underlying additional software or system dependencies necessary to run the new software or system you will create.

We do not foresee any software or system dependencies other than a server on which to run the software and the computational power (available, for example, through commercial providers such as Amazon Web Services, through institutional/academic processors, or through supercomputing centers).

B.4 Describe the processes you will use for development documentation and for maintaining and updating technical documentation for users of the software or system.
Documentation will be a required deliverable for all engineers and engineering teams working on the project. An example of the general format for documentation can be seen at https://github.com/APMG/audio-search and https://github.com/popuparchive/pop-up-archive.

WGBH Digital will produce detailed documentation for installing and implementing the crowdsourcing tool's source code. WGBH MLA developers will review the documentation to ensure that it is understandable to external users. All source code and documentation will be made available on the WGBH Github repository. WGBH digital uses Atlassian's Confluence for documentation, JIRA for task and bug tracking and Github for exposing the code.

B.5 Provide URL(s) for examples of any previous software tools or systems your organization has created.

WGBH:

- American Archive of Public Broadcasting: http://www.americanarchive.org
- Boston TV News Digital Library: http://www.bostonlocaltv.org
- Open Vault: http://openvault.wgbh.org
- Design Squad, recent winner of an Emmy award: http://pbskids.org/designsquad/
- News and Then, for American Experience, runs off an API created by the Public Media Partnership: http://www.pbs.org/wgbh/americanexperience/newsandthen/
- Antiques Roadshow, fully responsive site, also built an API for all of their appraisals/appraisers: http://www.pbs.org/wgbh/roadshow/
- Plum Landing: http://pbskids.org/plumlanding/
- Downtown Abbey, Season 5: http://www.pbs.org/wgbh/masterpiece/downtonabbey/downton-experience.html
- Forum Network, includes customized backend to house more than 5000 videos: http://forum-network.org/

Pop Up Archive:
https://github.com/APMG/audio-search

https://github.com/popuparchive/pop-up-archive

C. Access and Use

C.1 We expect applicants seeking federal funds for software or system development to develop and release these products as open source software. What ownership rights will your organization assert over the new software or system, and what conditions will you impose on the access and use of this product? Explain any terms of access and conditions of use, why these terms or conditions are justifiable, and how you will notify potential users of the software or system.
WGBH and Pop Up Archive will each own components of the digital content (code and research data) created through this grant. Code developed for speech-to-text, audio waveform analysis, and crowdsourced metadata games will all be developed from previous open-source efforts and final products will also be available and open-source.

We will restrict use of the code and research data to non-commercial applications.

C.2 Describe how you will make the software or system available to the public and/or its intended users.
The Project Team will be highly involved in outreach and dissemination of the audio analysis and crowdsourcing tools and accompanying documentation to potential users of the tools including researchers and other institutions that might benefit from installing the tools locally to engage in computational and crowdsourcing efforts to improve access to their audiovisual collections. The tools will be shared with potential users via listservs and social media, and a press release to major print publications, professional organization journals, and listservs. The Project Team will also propose panel presentations to share findings and results at professional library, archives, and educational conferences, such as AMIA, SAA, NEA, and PBS.

Part IV. Projects Creating Research Data

1. Summarize the intended purpose of the research, the type of data to be collected or generated, the method for collection or generation, the approximate dates or frequency when the data will be generated or collected, and the intended use of the data collected.

The research data generated by this grant will include initial and improved metadata for audiovisual material resulting from computational methods and crowdsourced metadata cleanup and contributions. It will include a public database of audiovisual metadata (audio fingerprint) of machine-generated transcripts, descriptive tags, and waveform attributes such as speaker identities and other qualitative data like applause or laughter. Evaluation of the project and research will take the form of surveys and usability studies which will be analyzed and shared with the community via project reports.

2. Does the proposed research activity require approval by any internal review panel or institutional review board (IRB)? If so, has the proposed research activity already been approved? If not, what is your plan for securing approval?
No.

3. Will you collect any personally identifiable information (PII) about individuals or proprietary information about organizations? If so, detail the specific steps you will take to protect such information while you prepare the research data files for public release (e.g. data anonymization, suppression of personally identifiable information, synthetic data).
No.

4. If you will collect additional documentation such as consent agreements along with the data, describe plans for preserving the documentation and ensuring that its relationship to the collected data is maintained.
N/A

5. What will you use to collect or generate the data? Provide details about any technical requirements or dependencies that would be necessary for understanding, retrieving, displaying, or processing the dataset(s).
We will use and develop open-source software (see Part III) to create machine-generated metadata as well as the platform for collecting crowdsourced corrections and improvements to the metadata. The software will be built from or consist of training for the open-source libraries Kaldi, ARLO, and Tilt Factor's Metadata Games. We will work with engineers familiar with all three libraries and will be advised by the teams supporting ARLO and Metadata Games. To evaluate the project, the team will use survey tools such as Survey Monkey. Results of usability studies will be documented by the project team. All evaluations will be shared as pdf reports and shared with the community on the AAPB website.

6. What documentation will you capture or create along with the dataset(s)? What standards or schema will you use? Where will the documentation be stored, and in what format(s)? **How will you permanently associate and manage the documentation with the dataset(s) it describes?**
Documentation will be stored publicly on the web in locations such as Github, the AAPB website, and disseminated among the open-source communities that are active in the open-source communities to which code will contribute (i.e. Kaldi, ARLO).

7. What is the plan for archiving, managing, and disseminating data after the completion of research activity?
The data will be made public and easily downloadable so it can be disseminated widely. In addition to public locations such as Github, master material and digital data will be kept in a secure, climate-controlled, fireproof vault; other materials are maintained within a nearby secure storage location.

8. Identify where you will be publicly depositing dataset(s):

Name of repositories: WGBH Github Repository, Pop Up Archive Github Repository
URL: https://github.com/WGBH, https://github.com/popuparchive

9. **When and how frequently will you review this data management plan? How will the implementation be monitored?**

The WGBH Media Library and Archive (MLA) are responsible for the safekeeping of all project data. They assure the present and future accessibility, sustainability, and security of project media and assets.

WGBH has established a retention and destruction policy for documents received or created by the Foundation to ensure compliance with legal requirements and to protect the Foundation's intellectual property. This policy encompasses all documents in all their forms, electronic and hard copy, throughout their lifecycle.

Financial records, supporting documents, statistical records, and all other records pertinent to a Federal Grant Award are retained for 8 years after the final report submission date. If there is any question of litigation, claim or audit, these records are retained indefinitely.

Original Preliminary Proposal

IMLS Research Grant: **Improving Access to Time-Based Media through Cataloging Tools**
WGBH Educational Foundation: American Archive of Public Broadcasting and Pop Up Archive
January 2015

Libraries and public media archives across the country house tens of millions of audio and video recordings,[1] of which spoken word recordings are a significant component. However, no nationally established methods exist for digitally indexing and analyzing the audio in order to enhance the descriptive data and improve discoverability. As leaders in digital preservation, access, and analysis of culturally significant audiovisual content, WGBH and Pop Up Archive are seeking $900,000 for a Research Grant to establish such methods through the American Archive of Public Broadcasting (AAPB). AAPB, a collaboration between WGBH and the Library of Congress, is coordinating a national effort to identify, preserve, and make accessible as much as possible a digital archive of public television and radio dating back to the late 1940s. The initial collection consists of 40,000 hours of digital media selected by more than 100 public media stations and organizations.

This project will fund 1) development and use of audio tools to create transcripts of the 40,000 hours of AAPB digital files; 2) development of open-source web-based tools to improve the transcripts and descriptive data by engaging the public in a crowd-sourced, participatory cataloging project; 3) measuring improved access to media files with added crowd-sourced data; 4) publishing initial and improved data sets to 'teach' tools; and 5) building a public database and set of audio analysis tools identifying speaker audio wave patterns (audio fingerprint).

Research question

The dilemma for media material is that content on obsolete formats needs to be digitized to be accessible, and described to be identified and discoverable. With no robust descriptive metadata for the media, having 40,000 hours of digital content available on the web will not make it discoverable. To improve discoverability and access for scholars and researchers, the content needs better descriptive information that can be indexed by search engines. The resources necessary to catalog 40,000 hours of programming are overwhelming and unattainable. For time-based media, the content needs to be seen or heard in real time for a human to be able to describe it. Harnessing the interest of the public and spreading the work over many volunteers may be a solution to the lack of resources available to describe this content. The challenge will be to make the tools for capturing the data easy, intuitive and engaging – tools that anyone could use.

A number of attributes and data could be gathered without having to fully watch or listen to a program, which would greatly enhance discoverability. In addition, emerging technology increasingly enables metadata to be created automatically for digital time-based media through programmatic analysis and speech-to-text software. Automatically generated metadata could provide a map of content and topics addressed in US public broadcast material from the past 50 years and an authoritative index of nationally significant speaker voices. We seek to augment 40,000 hours of U.S. public broadcast content from radio and television stations of all sizes and locations with crowd-sourced and automatically generated metadata to serve as a national "jumping off point" for digital discovery and analyses of US public media and archival spoken word audio.

Methodology

This project will leverage emerging technology to create large quantities of metadata for digitized audio-visual content through highly accurate speech-to-text software, semantic tagging, and digital waveform data. Through metadata games and tools, crowd-sourced participants will help verify and correct transcripts from speech to text tools, add additional data, and identify speakers to add to a national audio fingerprint database.

---

[1] In 2007, the Association of Research Libraries estimated that its 123 member libraries held more than 10 million audio recordings (National Recording Preservation Board,*The State of Recorded Sound Preservation in the United States: A National Legacy at Risk in the Digital Age*, http://www.clir.org/pubs/reports/pub148/pub148.pdf, p.10). A Heritage Health Index published in 2005 revealed that 46.4 million collection items of recorded sound were housed within the institutions that participated in a baseline cultural heritage study. See Heritage Preservation, Inc., *A Public Trust at Risk: The Heritage Health Index Report on the State of America's Collections*, 162.

Pop Up Archive has already identified and contributed to development of state-of-the-art speech-to-text technology from institutions such as the International Computer Science Institute in Berkeley, CA. This project will enable WGBH and Pop Up Archive to build on this fundamental technology by creating an authoritative national speaker database, a taxonomy for public media and culturally significant archival audio, and a comparative analysis of 40,000 hours of historic recorded sound. We are budgeting $400,000 for the creation of this National Archival Audio Fingerprint (transcription, tagging, and indexing of 40,000 hours/voices in addition to research and development of analysis tools). We are budgeting an additional $200,000 for program salaries related to the creation of the database and analysis tools, including web development for the final public repositories and access points. An additional $300,000 will build the crowd-sourced tagging tools and management of the participatory crowd-sourcing endeavor.

WGBH and Pop Up Archive plan to build and improve on the work of COMMA[2], a BBC R&D project to explore new technologies in automated metadata extraction. Specifically, we are interested in building upon the BBC's work to create a national speaker database, to compare automatically-extracted metadata to pre-existing taxonomies (i.e. DBPedia or, in the case of the BBC, the World Service Archive), and to compare historical recorded sound to contemporary news coverage and conversations or areas of influence in social networks. For the crowd-sourced component, we will build a platform that guides people through a series of steps to enable metadata tagging without requiring existing knowledge of metadata tagging. Tools, workflows, and methodology will be shared and useful for other similar endeavors. This project will test the viability of using the public and simple metadata gathering tools to add descriptive data to records of digitized media to increase access and discoverability.

Outputs and dissemination

The first phase will result in a database of time-stamped transcripts and keyword tags for the entirety of the AAPB 40,000 hours and metadata games for the public to enhance the automatically generated data. The second phase will result in the National Audio Archival Fingerprint: a public repository of nationally significant audio "voice signatures" and an open-source codebase of tools to enable other archives to conduct analyses of their audio collections. The speaker repository and tools will be published under a Creative Commons license as a stand-alone product in multiple locations, such as Github. We will work with an advisory board composed of leaders in national archival, library, broadcast, and digital stewardship efforts, including Emily Gore (Digital Public Library of America), Laura Soto-Barra (NPR), and Roger MacDonald (Internet Archive), to widely disseminate the products of this grant.

Together, WGBH and Pop Up Archive represent the pinnacle of public media service and the vanguard in digital audio innovation. Broadcasting since 1951, WGBH currently maintains an archive of over 750,000 items and the public website Open Vault. Karen Cariani, Director of the WGBH Media Library and Archives, will lead the project. Funded by the John S. and James L. Knight Foundation and National Endowment for the Humanities, Pop Up Archive has indexed almost 1,000,000 minutes of recorded sound since 2012 as part of its centralized repository and tools for analyzing public media and archival audio, in collaboration with the Public Radio Exchange and under the advice of the British Broadcasting Corporation R&D division. Anne Wootton will lead the Pop Up Archive team.

The 2014 National Digital Stewardship Agenda includes, "Engage and encourage relationships between private/commercial and heritage organizations to collaborate on the development of standards and workflows that will ensure long-term access to our recorded and moving image heritage."[3] These partnerships are critical in order to move the needle of audiovisual access issues of national significance. WGBH and Pop Up Archive are eager to continue building such a relationship so that the innovations in technology, workflows, and data analysis advanced by the private sector are fully and sustainably leveraged for U.S. public media and cultural heritage organizations.

---

[2] http://www.bbc.co.uk/rd/projects/comma

[3] http://www.digitalpreservation.gov/ndsa/documents/2015NationalAgenda.pdf, p. 20.