Lyrasis® | Catalyst Fund

Applicant Organization: Emory University-Woodruff Library	Amount: $28,047.00
Proposal Title: Increasing Accessibility of Audiovisual Content Using Whisper	
Project Goal: This project's goal is to assess the viability of the open-source AI software tool Whisper as an accurate and cost-effective solution to captioning and transcribing audiovisual content in library, archive, and museum collections, and to develop resources that can be widely used by members of the Lyrasis community to caption and transcribe digital audiovisual materials at large scale, increasing the discoverability, searchability, and accessibility of these materials for all audiences and users.	

Project Description:

1. Describe the problem, need, issue or challenge that your project will address, why your project is innovative, how it could help others or advance knowledge or practice for the field, and what potential it has to scale as a solution for others.

Many libraries and universities share the challenge of ensuring that digitized audiovisual (AV) content is accessible to all audiences. Captioning digitized AV content on a proactive basis improves accessibility for end users; supports diversity, equity, and inclusion; and brings content closer into compliance with Web Content Accessibility Guidelines. Providing transcripts of AV content in collections also improves searchability and discoverability, thereby increasing access to underrepresented voices and communities within collections.

However, providing captions is currently resource intensive. The rate for the lowest tier of service with 3Play Media, Emory's vendor for captioning and transcription of AV content, is $114 per content hour, translating to approximately $39,900 to caption the AV content that the library digitizes each year. Solutions by companies such as Amazon and Microsoft offer lower rates but introduce data privacy and content accuracy concerns.

Solutions employing automated speech recognition (ASR) can generate captions for free or at little cost but have so far not achieved adequate accuracy. Editing ASR-generated transcripts to achieve equitable access standards of accuracy often requires over 4 hours of effort per content hour. At this rate, approximately 1400 hours of staff time (.67 FTE) would be needed to produce accurate captions for the amount of AV content that Emory digitizes each year.

As a result, the AV content at Emory Libraries, like so many institutions, is mostly uncaptioned and does not have available transcripts. This limits the searchability and discoverability of this content for users, and it renders the auditory information in this content inaccessible to users who are deaf or have hearing loss. To address this challenge, we propose a project that aims to provide a model that can be adopted by academic libraries, special collections, research facilities, museums, and archives of all sizes. A low-cost solution that runs locally and generates highly accurate captions would advance the dual objectives of discoverability and accessibility of content while providing a scalable model for integrating captioning into workflows and addressing this challenge across the Lyrasis community.

2. Describe your project plan, including activities, timeframe, resource requirements and, if relevant, collaborators and sustainability plans.

We propose testing Whisper, a recently released open-source AI software tool, for its viability as a solution to generate captions and transcripts for library AV content. Whisper performs multilingual speech recognition and speech translation while producing time-synchronized caption files and plain-text transcripts. All computing is done locally; no content or information is uploaded to the cloud or sent to a vendor. Through the capabilities of this newly released software, our project complements and expands upon the findings of the 2020 University of Mississippi Catalyst Fund project, while attempting to address some of the challenges it encountered.

We ran Whisper on a small set of AV content and found that the processing time and accuracy of results suggest it could be a feasible solution for captioning AV content on a large scale. With appropriate hardware, Whisper processed content at a rate of 25-30 minutes per content hour; at this rate, first-pass captions could be generated for a year of Emory's digitized AV content in less than 8 days. The results also achieved an average word error rate (WER) of 1.06%. This is better than the WER achieved by any other ASR solution available and is comparable with the accuracy rates guaranteed by human transcription services, such as Rev and 3Play Media. Given its accuracy, technical simplicity, data privacy, and open source nature, Whisper has strong potential to meet the needs of a proactive library captioning workflow.

Our project will test Whisper across a larger, more varied sample of AV content, further assessing its accuracy and determining the feasibility of deploying it within an AV digitization workflow. Emphasizing community-driven collections similar to collections held by many Lyrasis members, we intend to caption and transcribe approximately 250 hours of content, evaluating Whisper's overall performance and assessing its equity of performance for content from underrepresented communities. Content will be chosen to represent a multitude of contexts including technical and subject-specific vocabularies, multi-language material, regional dialects and vernaculars, multi-speaker content, environmental noise, and a range of sound and production qualities. Anticipated deliverables include hardware and software configuration recommendations; a standardized workflow to generate, process, edit, and deliver captions and transcripts; a style guide with subject-specific guidelines to standardize editing and formatting, incorporating consultation with accessibility experts and best practices for captioning and transcription; and establishment of benchmarks for assessing processing time, cost, accuracy rates, and editing time.

Funds will be used to hire student transcribers to edit generated caption files, to consult with accessibility and subject matter experts in the creation of the style guide, and to set up a platform (we anticipate Aviary) to test the end-user discoverability, searchability, and synchronization of captioned and transcribed content. Emory's in-kind contributions would constitute hardware to run Whisper and staff time to administer the project and produce its deliverables. If successful, future directions for research could include expanding the range of subject-specific guidelines or building additional Whisper functionality and user-friendly features for institutions with fewer technical resources.

In summary, this project's outcomes could provide a solution to a common challenge among the Lyrasis community. If Whisper achieves accuracy rates comparable to those of commercial captioning services, it could eliminate the data privacy concerns of using vendors and reduce captioning costs enough to incorporate captioning into AV digitization workflows, make digitized AV content accessible and text-searchable on a proactive basis, and provide resources and a scalable model for large-scale captioning and transcription of AV content across the Lyrasis community.

Projects involving OSS must *briefly* address sustainability, including governance, resources, community engagement, and technology as described in the ITAV framework at https://www.lyrasis.org/programs/Pages/IMLS-OSS.aspx

Released to Github under the MIT License on Sept. 21, 2022, Whisper is in Phase 1 for all four ITAV facets. There were 11 active authors and 23 active commits to the main branch in January 2023; there is a robust user community and extensions, integrations, and ports of Whisper are encouraged. The sustainability of Whisper's open-source release is critical only in terms of performing speech to text transcription safely, affordably, and at scale; its output formats (.srt, .vtt, and .txt files) are broadly available and well supported by speech to text vendors and media player interfaces.

3. Provide names and titles for the principal investigator(s) and other key participants in the project.

Principal investigator: Nina Rao, Audiovisual Conservator
Key personnel: Kyle Fenton, Head of Digitization Services; Simon O'Riordan, Head of Metadata Services
Other personnel: Andrew Battelini, Metadata Analyst; Jonathan Coulis, Oral History Coordinator

Budget.

Line	Basis	Cost
Student transcribers - undergraduate	840 hours at $16/hour wages + 7.65% benefits	$14,468.00
Student transcribers - graduate	420 hours at $20/hour wages + 27.8% benefits	$10,735.00
End-user test platform (Aviary)	12 months at $99.95/month	$1,200.00
File hosting for test platform	3.4 TB for 12 months at $5.99/TB/month	$244.00
Subject consultants for style guide	20 hours at $70.00/hour	$1,400.00
Total Budget Request		$28,047.00