AI & Machine Learning Project FAQ

Project: Audio to Text

As part of the Ontario Council of University Libraries (OCUL) AI and Machine Learning Initiative, this project evaluates Whisper, an audio-to-text AI model, to improve digital accessibility through the automated generation of transcripts for audio and audiovisual materials in academic libraries. For this purpose, the project examines the feasibility of hosting AI tools centrally on Scholars Portal infrastructure and the quality of generated transcripts.

February 2026

What is an audio-to-text AI technology?

Audio-to-text technology, also called speech-to-text, voice-to-text, or automatic speech recognition, is an artificial intelligence tool that recognizes spoken audio and converts it into written text.

What is the goal of the Audio to Text project?

This project seeks to establish a consortially-hosted workflow for Whisper processing, create testing and technical documentation for reuse, and report to OCUL member libraries on the testing of Whisper for audio file transcription.

How is testing conducted and what audio is included?

A small test corpus consisting of materials from the University of Toronto Archives and Records Management Services (UTARMS) and appropriately licensed Open Educational Resources (OER) were used for the project. These materials included a range of lengths and formats to simulate the variety of audio and audiovisual materials found in member library collections. Generated transcripts are being evaluated for quality by a group of member volunteers from the OCUL Video Community and OCUL Accessibility Community using a standard rubric developed and validated for the project.

How are intellectual property, copyright, and privacy being considered and preserved?

Copyright for materials from UTARMS used in this project is held by the university and used with permission. OER included in the test corpus were selected intentionally for reuse that allowed non-commercial adaptations. The instance of Whisper being used for this project is in a closed, ‘walled garden’ environment where materials are processed on centrally-supported Scholars Portal infrastructure.

What has been learned so far from the project?

We have learned that it is possible to set up a functional workflow for processing files on Scholars Portal infrastructure, but only after purchasing upgraded hardware to run the larger Whisper model required for this project. We also learned a great deal about how to evaluate transcripts for quality and what licensing and copyright considerations are present for audio and audiovisual materials.

Our work with Whisper provided insight into both the capabilities and the limitations of the speech to text model when applied in libraries. Volunteers who reviewed a corpus of test files agreed that the overall quality of the transcripts produced by Whisper exceeded their expectations. However, while the model demonstrates a baseline level of performance and support for a range of languages, we noted some limitations.

We found that out-of-the-box Whisper settings require additional parameter tuning to improve the quality of transcripts. Common issues we encountered include missing punctuation, inconsistent capitalization and line breaks, hallucinations where the model added content not part of the original recording, and language issues for non-English audiovisual content. Additionally, Whisper does not denote or differentiate between speakers. For multi-speaker recordings, speaker labels need to be added manually to appropriately capture and represent narration and dialogue. If the transcript is used for captions, some manual review and alignment is typically still needed. The feasibility of this manual review depends on how much of the transcript requires remediation as well as how these changes will impact timestamps.

Overall, our refinements to the Whisper model script revealed the quality of transcriptions is dependent on the ability to fine-tune parameters and iterative experimentation. Improving transcript quality can also mean higher resource use or longer processing time, so there is often a consideration of the trade-offs between transcript quality, speed, and computational cost.

What aspects of this project might be useful to OCUL member libraries?

While this project focuses on establishing and testing speech-to-text processing with Whisper, if successful, the project could result in a new service available to OCUL member libraries. Even if a new service is not established, the reports, documentation, and validated evaluation rubric for speech-to-text transcripts are useful for any member library who wishes to consider a speech-to-text workflow at their library.

Have questions or project feedback?

Reach out to Program Manager Kari D. Weaver: kari.weaver@ocul.on.ca.

OCUL
Ontario Council of University Libraries

ocul.on.ca/aiml-program