

Out of the Archives

a newsletter from the Washington State Archives

July 2016



Accessibility via Technology The Holy Grail of Audio Recognition

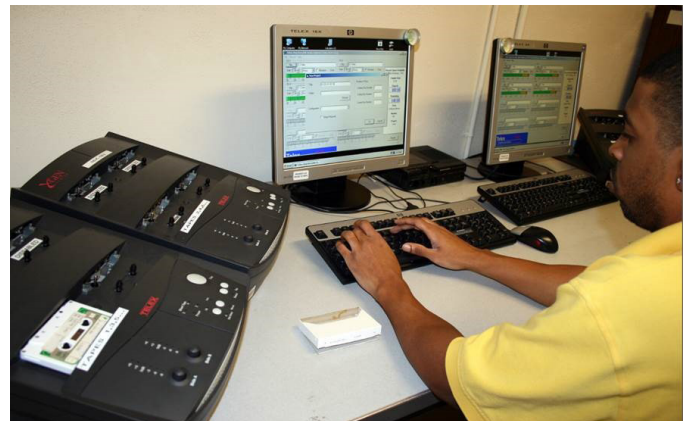
by Larry Cebula, Assistant Digital Archivist

Big news today from the Washington State Digital Archives! (Full disclosure: I am an Assistant Digital Archivist here.) Today we put the audio files of the House of Representatives Committee Meeting Recordings online--and they are keyword searchable.

The House of Representatives Committee Meeting Recordings cover 1973 to 2001. This is almost 6000 hours of hearings and the files take up 1 terabyte of data.

This list of house committees will help give context to some of these files. The files came from 30,000 cassette tapes. The tapes were converted to digital files and cleaned up starting in 2005. Putting them online and making them searchable is a cooperative project between the Washington State Digital Archives and the Microsoft Corporation.

The technical breakthrough is that these files are keyword searchable. Users can enter keywords or phrases and the search engine will dig through all of the files and discover when anyone spoke those words. The search results give some details about the file but also a snippet of the text showing where on that file the words were spoken. Click on any of the strings of words between the dashes and the in-line player will take you directly to that point in the recording. Some good keyword searches



are salmon and dams, "Indian gaming," "state history," and "Lewis and Clark."

This, my friends, is one of the holy grails of computing: untrained voice recognition over thousands of hours of tapes and many different voices. We rolled out this technology with the legislative hearings because we are a state archives and this gives the Washington State public unprecedented access to these public records. But think of the other uses for the keyword searching of audio files. I have never visited an archives that did not have boxes of decaying audio tapes from an oral history project that never quite got to the transcribing stage. These tapes can be digitally preserved and put online. Television and radio interviews and news and talk programs will become searchable. This is a digital history breakthrough.