Making Public Records Legible: An Open-Source Platform for Transcribing Archived Multimedia

Nicholas Weber[1] and Jackson Brown[1]

The Information School, University of Washington
nmweber@uw.edu

Abstract. In the following paper we describe an open-source platform that includes a transcription engine and an interface for crowdsourced editing of legislative documents. We present an implementation of this platform for the City of Seattle, and situate this work in the context of research infrastructures that support public humanities research.

Keywords: Transcription, Civic Technology, Public Humanities

1 Public Records Archives

"It will be of little avail to the people that the laws are made by men of their own choice if the laws be so voluminous that they cannot be read, or so incoherent that they cannot be understood." James Madison 1788

Government transparency is a democratic ideal often championed by elected officials in the name of accountability. Recent initiatives focused on the publication of open data, and the development of civic technologies are illustrative examples of how these ideals are being mapped onto a contemporary information communication technology (ICT) landscape. Unfortunately, the archives produced by these well-intentioned initiatives (e.g. open data portals, legislative document repositories, etc) often obscure as much as they reveal [2] . That is, these initiatives often produce archives of *public* records that are difficult for the general *public* to search, access, and use. The practice of preserving and making accessible the records of a government - in a form which the polity is equipped to meaningfully use - is a traditional dilemma for archivists and historians. However, we argue that as public records are rapidly produced in the form of complex digital objects there is a need to reconsider traditional archival approaches to preserving and making these artifacts accessible.

In our current research, we have developed a web-based application to transform media recordings (audio and video) into text transcripts, an indexing service for resolving queries against these transcripts, and an interactive web-interface for crowdsourced editing of individual transcripts. The open-sourced transcription engine we have developed can be adapted to both web-based scraping of websites that link to multimedia archives, as well as an API plug-in for one

of the most commonly used content-management systems for legislative record-keeping. The following sections describe a use case, the software dependencies, and the system architecture of this application.

1.1 Public Records of Local Governments

City councils in the USA often record legislative sessions as digital video. The video-recordings are extremely valuable in that that they exhaustively capture the proceedings of a public legislative session (e.g. Who speaks, when, what is said, in what context, etc.). However, these recordings are archived as the *only* public record of a legislative debate. Without a text-based transcript the content of these recordings are broadly inaccessible for contemporary and historical scholarship. For example, imagine a public administration scholar wanting to understand the rhetorical frames used by incumbent City Council representatives over the last decade. To undertake this research the scholar would need to search ten years of relevant videos (based solely on the title of the legislative session), watch each video to determine its relevance to her research questions, and then manually transcribe the speech acts of individual City Council members. This mode of access is incredibly labor intensive, and unlikely to result in reliable historical scholarship. So, while the archived video-recordings of legislative debate are exhaustive public records, in their current form they are of limited value to the general public.

1.2 Platform Architecture

We have developed a web-application using a number of open-source tools to help remedy this problem. Below, we briefly sketch the main components of our application for transforming multi-media records into editable transcripts.

- Video Files: We first obtain video files by either scraping a website where links are posted, or by querying an API to a content management system in use by city council's throughout the USA (Legistar).
- Video to Audio: We then transform each video to an audio stream using the 'FFMpeg' framework [1].
- Audio to Text: Using the Python package SpeechRecognition [5] we separate each audio file into discrete seventeen second segments, and then call the Google Speech Recognition API - a free service for transforming audio streams to digital text - to transcribe each audio segment. Our transcription engine then recombines each seventeen second chunk into a complete text transcript.
- Corpus Indexing: For all transcribed legislative sessions, we implement fuzzy search using the information retrieval approach Term Frequency - Inverse Document Frequency (TF-IDF). We construct a tree of information about each transcribed document in the corpus, and then index this information for the entire corpus. After creating the tree, we implement a simple searching method that takes a search phrase (provided by the user of our application)

and read the tree to find the most relevant nodes (documents) by their weighted TFIDF score.
- GUI: The web-application then publishes the transcript and video to the web at permalinked page. The web-application also includes an interface for users to search for, discover, and edit transcripts that may have errors (due to the automated transcription provided by the Google Speech Recognition service).

1.3 Implementation

We have developed a single implementation of this web-application for the City of Seattle Washington available at https://councildataproject.github.io/seattle/. This implementation is a proof of concept for public testing. In future work, we will describe evaluation with end-users, and the development of additional features for transcript editing, anchoring video to transcript sections, downloadable transcripts, and automatically recognizing turn-taking of speakers.

1.4 Conclusion

Public scholarship - the "making knowledge 'about, for, and with' diverse publics and communities" [4] - has been a common thread amongst humanists focused on the development of public records archives. The project we have described is aimed at transforming the documentary practices of state and local governments from multi-media recordings to discoverable and searchable textual archives. Through this process we argue that our application makes public records legible for the public, improves the transparency of local governments, and enables public scholars to use text-based corpora for answering research questions about the history of local governance.

References

1. Bellard, F., & Niedermayer, M. (2012). FFmpeg. Available from: http://ffmpeg.org.
2. Janssen, M., Charalabidis, Y., & Zuiderwijk, A. (2012). Benefits, adoption barriers and myths of open data and open government. Information systems management, 29(4), 258-268.
3. Madison, J. (1788). The Federalist Papers, No. 62. Independent Journal, February, 27.
4. Woodward, K. (2009). The future of the humanities-in the present & in public. Daedalus, 138(1), 110-123.
5. Zhang, A. (2017). Speech Recognition. Version 3.7. Available from: https://github.com/Uberi/speechrecognition