Enhancing ICRC sound recordings metadata using NLP

**Laboratory:** EPFL Swiss Data Science Center

**Type:** Semester Project

**Problem statement**

The ICRC sound collection counts approximately 8400 sound recordings. 6400 of those files were generated by a digitization project, 2000 are born digital. They represent a total of approx. 2800 hours of sound. The contents of the recordings are various: ICRC broadcast productions, conferences, speeches, interviews, music, spots; in several languages. Each recording is manually described and indexed in a database. This task is very time-consuming, and the metadata added by the archivists are minimal. At a time when audio mediums become increasingly popular, the Archives acquire very few new recordings due to a lack of resources.

Since 2016, the ICRC Archives have an online audiovisual portal, with a public and an internal interface. Lack of comprehensive metadata and robust description stand in the way of access, ultimately resulting in the inability to derive full value from digitized and born-digital collection of audio content.

**Goals of the project**

- To enhance the discoverability and usage of the sound collection
- To enhance the collections with new recordings
- To facilitate access to the documents content, including for persons with hearing disabilities

by enhancing the recording metadata, specifically:

- Using Automatic Speech recognition (Speech-to-text) to generate transcripts, with timecodes.
- Entity extractions /indexing

**Datasets/record samples that can be made available for the purpose of this project:**

- 148 files “Oral history” representing 108 hours of recordings, Language: French
- 148 files “debriefing” representing 160 hours of recordings, Language: English or French
- 174 files “public communication” representing 75 hours of recordings, Language: English or French

Further info: total size of the files is 19,06 GB compressed, format of the files: mp3

**Expected project deliverables:**

- Transcripts
- Algorithm

**Contact:** Roberto Castello (roberto.castello@epfl.ch)