

Data Management Plan

CITF-Postdoc

Zack Batist

2025-01-15

Note

This document is still a work in progress.

This document describes the data that this research will produce and the procedures for curating data throughout the project.

Kinds of data

This project relies on various data sources, including interviews, bibliographic resources and structured datasets.

Interviews generate audio, video and textual resources, formatted according to the WAV, MP4 and markdown specifications, respectively.

While reviewing extant literature, I compile numerous published PDF and HTML documents. I maintain a bibliographic database using [Zotero](#) and continually export core information to [BibLaTeX](#) format using the [Better BibTeX](#) add-on.

I prioritize writing all documents using plaintext file formats. I rely on [Quarto](#), an open-source scientific and technical publishing system, to generate HTML, PDF and DOCX files from singular markdown files via the [pandoc](#) conversion engine and the [LaTeX](#) typesetting system.

When collecting structured data, I opt to use open text formats including CSV, JSON and XML, however I may also use XLSX format to maintain interoperability with collaborators.

I embed all original code, including database retrieval queries and API requests, in or alongside quarto documents, where I annotate code with detailed and contextualizing comments.

I am still exploring my options for QDA tooling. If I decide to use [MaxQDA](#), the data will be stored in a project file comprising a proprietary SQLite Database schema. Alternatively, I may use [qc](#) which stores data across a series of text files and a barebones SQLite database.

Activities, processes and workflows

Data collection

Interviews generate audio, video and textual records. I primarily rely on a SONY ICD-UX560 audio recorder, which contains 4GB of internal storage supplemented with a 32GB microSD card.¹ I record using the lossless 16bit 44.1 kHz Linear PCM wav format.

With participants' consent, I also record video using a GoPro Hero 4 Silver action camera equipped with a 64GB microSD card.

I maintain typed notes before, during and after each interview, which I may edit and re-organize to enhance clarity.

Immediately after each interview, I copy the data off the recording devices and onto a dedicated project drive. I organize and rename files using a semantic naming scheme and then mirror all files onto physical and a cloud-based backup drives.

Processing data

I use [FFmpeg](#) and [Audacity](#) to cut, concatenate, and clean the audio and video files if necessary. I generate transcripts of audio recordings following the [transcription protocol](#).

¹ See <https://weloty.com/sony-icd-ux560-review> and <https://weloty.com/using-the-sony-icd-ux560-the-4-how-tos> for more information on this audio recording device.

Storage and backups

I keep all data on a dedicated portable solid state drive which serves as a working directory for all research activities. I mirror the contents of this drive onto an identical secondary backup drive and onto cloud storage administered by the CITF using [rsync](#).

I maintain this website where I share documentation that supports this project and reflect on the work as it progresses. It is hosted using GitHub Pages and is backed up using Dropbox, however no sensitive research data will pass through these services.

Publishing and archiving

After my research is complete, I will make a concerted effort to document all the data I will have collected and generated and deposit them in a digital repository operated by a dedicated team of digital archivists certified to curate and preserve research data in perpetuity.