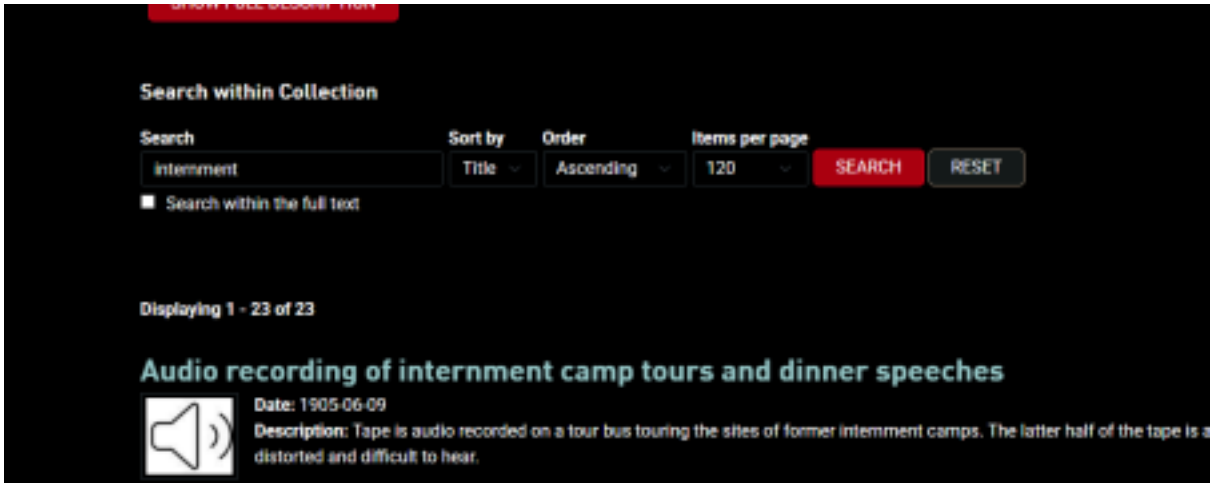
***Report 1: Capta and Mining***

J. Drucker (2021) states that data is not an entity that lives in the wild, waiting for a lucky hunter to come across and capture it. Instead, Drucker (2014) suggests that humans fabricate data to make sense of the world surrounding them, referring to this data as *capta* (p. 128). The term data, on the other hand, is something that is given, therefore, "able to be recorded and observed" (p. 2). Thus, capta has to be created, with this creation taking place through a "process of abstraction [...] called data modelling" (p. 19).

**Figure 1**
A screenshot of the data used for this report.
*Note*. Screenshot.




The Japanese Canadian Oral History Collection (Figure 1) comprises a diverse range of interviews with Japanese Canadians on various topics, touching on issues of "early immigration; their participation in pre-World War II industries including fishing, farming, and the lumber industries; and their internment during World War II" (Japanese Canadian Oral History Collection, 2025, p. 1). For this report, the tentative research question is: "How have Japanese Canadians experienced the internment?".

These interviews, saved as audio files in MP3 format, serve as an example of unstructured data, which, unlike structured ones, as Drucker (2021) describes, is not "composed of entities that are explicit, discrete, and unambiguous—like numbers or true/false statements" (p.19). Instead, unstructured data is "like natural language, is sometimes ambiguous and unclear" (p. 19). One of the biggest challenges in handling unstructured data, as in the case of literary text, is ambiguity, states Stephen Marche (2012), because "it contains a furiously distressed joy that words mean so much more than they mean" (p.6). In a similar vein, Hardy et al (2015) maintain that audio files are challenging because " the variability of audio motifs makes pattern mining difficult, [...] since the variability due to different speakers and channels is high " (p. 1).

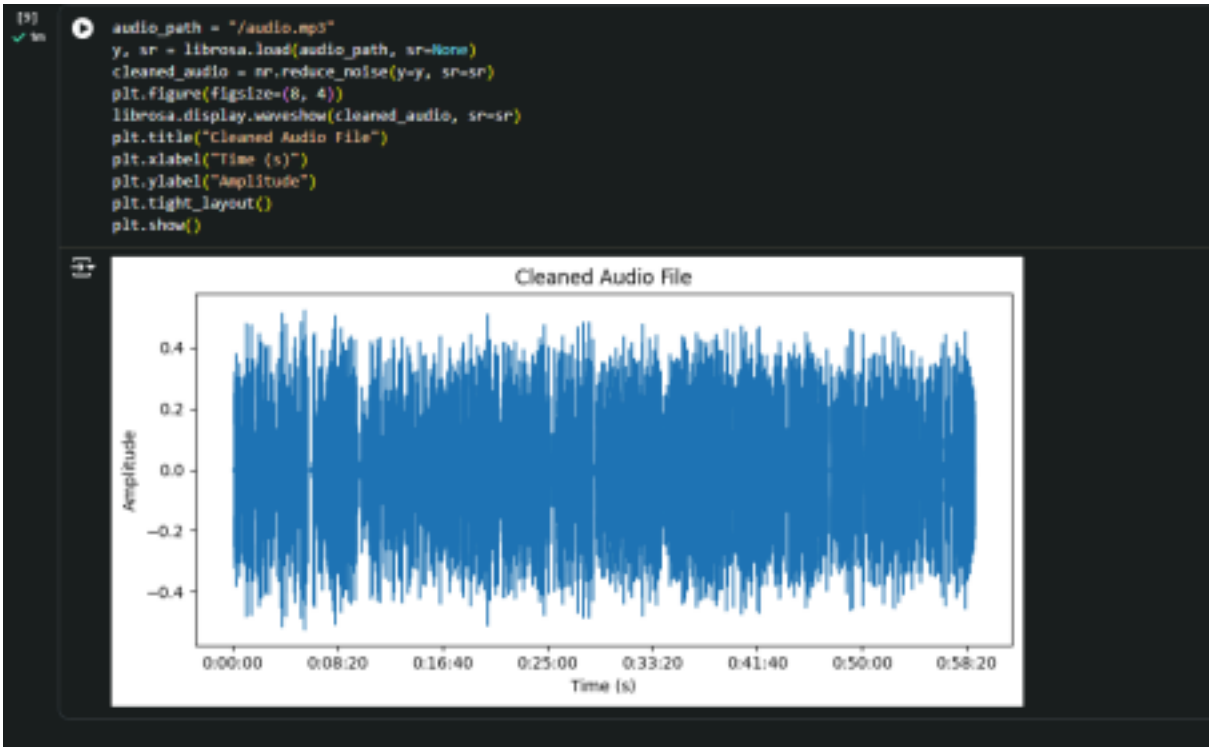
Drucker (2021) states that data mining has become an everyday staple in the natural and social sciences, used in"research methods in text, music, sound recording, images, and multimodal

communications studies with tools customi[s]ed for these purposes" (p. 110). Furthermore, Jockers and Underwood (2016) explain that data mining is an "informal name for a subfield of computer science [...] known as knowledge discovery in databases (KDD)" (p. 292). For Han et al. (2018), "[d]ata mining is the process of discovering interesting patterns and knowledge from large amounts of data" (p. 8).

For a preliminary analysis of the interviews, the first step is to select a specific range of data, which involves choosing audio files focused on the topic of Japanese internment. Specifically, by using the search field available on the website, the term 'internment' is entered, which returns 120 relevant files (Figure 1).

The second step, as Zilliz (2025) recommends, is preprocessing, which means "transform[ing] raw, unstructured audio data into clean, usable, and standardis[s]d formats for analysis and model training" (p. 1). Data cleaning for audio involves omitting background noise, which can be accomplished using the librosa library (p. 1). Figure 2 shows how the file has been processed and plotted using some Python code.

**Figure 2**
Some preliminary data mining of one audio file.
*Note*. The graph is mine and my work alone.




(500 words.)

**Acknowledgements**

Apart from using Grammarly for sanity checking regarding glaring errors, no other AI tools were

used.

**References**


Drucker, J. (2011). Humanities approaches to graphical display. *Digital Humanities Quarterly, 5*(1).

—. (2021). *The digital humanities coursebook: An introduction to digital methods for research and scholarship* (1st ed.). Routledge. https://doi.org/10.4324/9781003106531

—. (2014). *Graphesis: visual forms of knowledge production*. Harvard University Press.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining: Concepts and techniques* (3rd ed.). Morgan Kaufmann.

Hardy, C., Amsaleg, L., Gravier, G., Malinowski, S., & Quiniou, R. (2015). Sequential pattern mining on multimedia data. *Proceedings 1st International Workshop on Advanced Analytics and Learning on Temporal Data*, *AALTD*, 64-70. https://doi.org/10.48550/arXiv.2302.01932

*Japanese Canadian Oral History Collection*. (2025, September). Library Digital Collections. SFU, Simon Fraser University. https://digital.lib.sfu.ca/japanese-cdn-audio

Jockers, M. L., & Underwood, T. (2015). Text‐Mining the Humanities. In S. Schreibman, R. Siemens, & J. Unsworth (Eds.), *A new companion to digital humanities* (pp. 291-306). Wiley-Blackwell.

Marche, S. (2012, October 28). *Literature is not data: Against digital humanities*. Los Angeles Review of Books.

Zilliz (2025, February). *Getting Started with Audio Data: Processing Techniques and Key Challenge*s. Medium. https://medium.com/@zilliz_learn/getting-started-with-audio-data-processing-techniques-and key-challenges-420dc5233163