Ruth Nordquist

LIS 545B

Final Report

February 23rd, 2024

“To bee or not to bee: An annotated dataset for beehive sound recognition”

I. Data and Metadata Profile (https://zenodo.org/records/1321278)

The data in this dataset include audio recordings of various beehives, accompanied by text annotations for each audio file indicating at which points the audible noise in a recording is being produced by bee activity, and at which points non-bee background noise can be heard (traffic, birds, etc.). These data originated from two sources. The data collected as part of the NU-Hive project consists of recordings of two beehives in controlled environments (the intent being to minimize extraneous variables for the purpose of studying bee behavior when the queen is present vs. when the queen is not present). However, as part of the Open Source Beehive (OSBH) project, members of the public voluntarily provided amateur recordings of their beehives alongside an identification of the state of the hive at the time of recording (active or inactive), and these recordings/annotations from 6 beehives represent the remainder of the entries in the dataset. These data providers (participants in the OSBH project and researchers working on the NU-Hive project) are among the key stakeholders for the data, along with the primary facilitators of the data collection and corresponding research initiatives (Inês Nolasco and Emmanouil Benetos). Other researchers whose areas of study overlap with the data collected on bee behavior in this dataset (or whose work cites it directly) are other potential stakeholders, as well as any parties which may have provided funding (though none are referenced in the dataset or accompanying metadata).

The dataset consists of 136 data files representing 78 total recordings of beehives and associated annotations (along with several master files of the collected clips or annotations, as well as plaintext documentation of metadata in PDF format). The audio files contain sound from each beehive recording, while the associated annotation files contain lines of text which log whether the noise heard during each span of the audio file is attributed to the bees (Bee) or to background noise (noBee). The audio recordings are primarily uploaded as .WAV files (with the exception of two MP3 files), while the annotations are largely held in .LAB files (the master annotation file being the exception in this case, since it is an .MLF file). I was able to access the basic text/audio from all files in this dataset, although it seems that several file formats (such as .lab) were intended to be opened with software that I do not have downloaded, so I'm unable to determine whether the text files I was able to access were altered or compromised in any way from their original intended format. Since all of the data in this dataset are licensed under a Creative Commons Attribution 4.0 International License, users are permitted to freely share and adapt the data so long as they "give appropriate credit, provide a link to the license, and indicate if changes were made" (Creative Commons, n.d.).

The metadata associated with this dataset includes creators, a creation date, a thorough description of the data (including documentation on the origin and structure of the data as well as associated projects/publications), a persistent identifier for the dataset, publisher, license information, and a title for the dataset. A plaintext version of the majority of this metadata is included in PDF format as part of the dataset, but you can also export the metadata separately in a variety of formats (including JSON and XML).

If we apply the principles associated with IFLA FRBR user tasks (IFLA, n.d.), I would say that the metadata is reasonably comprehensive in that it sufficiently supports the tasks of identifying and selecting the dataset. The description section in particular provides information which would allow a user to determine the subject, origin, purpose, associated research, etc. of the dataset, but I think the metadata could stand to include more details which would facilitate finding/discovery. I am still in the process of learning to identify JSON or XML formats (and to decipher indications within the file that a particular metadata standard being used), so the only indication I found in this case of the metadata being structured according to a standard or schema was the option to download the metadata in various structures/formats (Dublin Core XML file, DataCite JSON file, etc).

The lack of metadata relating to the subject of the dataset combined with the limited options for advanced searching in Zenodo makes this dataset very difficult to discover without prior knowledge of specific information to search for in quotes. Although there's not much one can do about the search functionality of Zenodo, I do think that the metadata for this dataset could be improved by the inclusion of subject tags such as "Insecta," since there are at least subject facets provided for narrowing your search by subject in Zenodo. In terms of obtaining and using the data once found, I'd say that the metadata is sufficient to broadly support these tasks.

The metadata for this dataset includes three references, the most notable being the formal writeup of the research conducted by the credited creators of this dataset, first presented at a DCASE workshop in 2018 (Nolasco & Benetos, 2021). The metadata also references the OSBH project (OSBeehives, n.d.) from which recordings of beehives were sourced for this dataset, along with one other conference paper presented in 2018 on sound emitted by bees (Cecchi et al., 2018). Finally, there is also one entry that appears to have been identified by Zenodo as a paper citing this dataset in its discussion of automated data collection and the digitization of beekeeping practices (Hadjur et al., 2022). This is in line with many of the other results that were returned from my own preliminary search in Google Scholar for entries flagged as citing "To bee or not to bee," which largely centered around discussions of beehive acoustics and the development of automated data collection and transfer methods for bee data. I identified the bulk of these resources by navigating from the Google Scholar entry for Nolasco & Benetos' article itself, although there were also numerous results returned for searches in other portals for “To bee or not to bee: Investigating machine learning approaches for beehive sound recognition.”

II. Repository Profile (https://archive.ics.uci.edu/)

I chose the UC Irvine Machine Learning Repository for “To bee or not to bee” (Nolasco & Benetos, 2018) because it felt like the most appropriate repository to support the original intent of the dataset. Specifically, it was one of the few repositories I was able to find which was intended to house datasets that were assembled for the purpose of training algorithms to accurately classify instances of data based on specific variables. There were a number of repositories for data relating to life sciences (and zoology), but in light of the fact that my chosen dataset was originally assembled with the intent to develop an automated means of recognizing and classifying beehive noise, it seemed to me that the most appropriate choice would be a repository which would support the data being used for machine learning.

The UCI Machine Learning Repository houses a donation-based collection of datasets contributed by a broad variety of data creators and donors, and apart from a few stipulations for donating a dataset (must have permission from the original dataset collector in order to donate, any Personally Identifiable Information must be removed from dataset before donating, DOI will be assigned if the dataset does not already have a DOI, and dataset will be under CC BY 4.0 license) (*Donation - UCI Machine Learning Repository, n.d.*), it seems as though any dataset which is deemed appropriate for the purpose at hand (as judged by the donation approval process) may be added to the collection.

According to the repository details recorded on re3data.org, accepted content types for submissions include standard office documents, archived data, plain text, and databases. The website for the UCI repository itself does not specify what the limitations are around accepted data types (beyond noting that it is a repository for “databases, domain theories, and data generators” (*About - UCI Machine Learning Repository, n.d.*), but when browsing the repository, one is able to use facets to filter according to 8 data type categories (Image, Multivariate, Sequential, Spatiotemporal, Tabular, Text, Time-Series, and Other). As far as I was able to find, there are no explicit limits on file format or on what domains are acceptable for dataset submissions, and content found within the repository varies broadly on both of these points.

Unfortunately, actually donating a dataset to the repository involves creating an account that is associated with your institution in order to register with UCI. Without having an account and actually initiating the process of donating a dataset, very few details are provided on the website itself regarding what would be included in the SIP. On the Donation Policy page for the repository, an email address is provided for any questions about the donation process, but that is the only human touchpoint provided to a potential submitter prior to their registering with UCI.

With regard to the metadata requirements, the profile on re3data.org lists the metadata standard as a case of “Repository-Developed Metadata Schemas” from the DDC. Based on the description of this metadata “standard” from the DDC website, this seems to mean that established standards were not a good fit for the repository, and that UCI therefore created their own requirements for metadata associated with their datasets (*Repository-Developed Metadata Schemas* | *DCC, n.d.*). As such, I was not able to determine whether the repository has any particular requirements around metadata included in the submission package for datasets.

Although a login is required to donate a dataset, it is not required in order to download data from this repository. Direct download is the primary mechanism provided for accessing data, but for some datasets, there is also an option to import in Python found directly under the download button. Before I discuss the more ambiguous elements of metadata requirements for the DIP, it should also be noted that one of the unique features of the DCI Machine Learning Repository is that each dataset entry includes a chart which presents all of the relevant variables for the given data. This would be particularly useful for the beehive audio dataset, since the entry would prominently display and describe each variable being notated in the annotations for each audio clip (bee or no bee, active or inactive, etc), making it clear to the user what their automated recognition system should be looking to identify within the recordings.

As far as the metadata standards are concerned, though, I have not been able to determine that there is an established standard in place for the metadata being displayed. There is no evident option for accessing a standardized metadata file for datasets in the repository, and the data downloads themselves are very spotty and unregulated when it comes to the inclusion, construction, and naming of metadata files. In fact, I was able to find very little consistency whatsoever when it came to what was included in the DIP for this repository. Each downloaded dataset was compressed into a zip file, but beyond that, I was not able to identify a consistent pattern for how the data files within were organized or constructed from one dataset to the next. As such, I felt I was only able to make suppositions about the UCI's personalized standards for metadata (or DIPs) based on the attributes and data formats which most frequently appeared among the metadata displayed for a given entry (or found among the downloaded dataset files).

III. Recommended data citation

My recommendation would be to use the archival data citation generated by Zenodo, which is associated with the persistent DOI that was assigned for this dataset:

Nolasco, I., & Benetos, E. (2018). To bee or not to bee: An annotated dataset for beehive sound recognition [Data set]. Zenodo. https://doi.org/10.5281/zenodo.1321278

IV. Considerations for long-term preservation

With regard to the various factors relevant to the long-term preservation of this dataset, it is worth considering how this data may be used in the future, what formats are represented among the data, whether the dataset is uniquely identifiable, the comprehensiveness of the associated metadata, to what extent backups may be needed, and possible complications relating to the size of the dataset (Hart et.al., 2016).

In the interest of mitigating potential future complications relating to how the data may be used in unexpected ways in the future, data should ideally remain in its raw format. It may also be desirable to convert existing .mlf, .lab, and .wav files into non-proprietary formats that will not require specific software to access, in the case that the software that the current formats rely on become obsolete in the future.

The dataset has the advantage of an assigned DOI that makes it uniquely identifiable, but with regard to metadata, both the longevity and discoverability of the dataset could be reinforced by expanding the existing metadata to be more comprehensive, particularly in its representation of key subjects for the dataset. This dataset also already has a backup in place since the data is already stored on Zenodo, so the final primary factor to consider which may impede preservation of the dataset long-term is the issue of large file size. The beehive recordings could not be uploaded to GitHub at all due to file size, and even repositories that do have the capacity to accept the full 4 GB of data may face future difficulties relating to data transfer, so this may prove to be a complication for long-term storage of the many 25 MB+ audio recordings that are a part of this dataset.

V. Copyright license statement

In Zenodo, this dataset is under the Creative Commons Attribution 4.0 International License (Nolasco & Benetos, 2018). This license “allows re-distribution and re-use of a licensed work on the condition that the creator is appropriately credited” (2018), which would be also appropriate for the purposes of the dataset within this repository.

VI. Statement on human subject considerations

Since the subjects of this dataset are bees rather than humans, and since personally identifiable data about individual contributors of beehive audio recordings are not included anywhere in the repository, no further steps are needed in order to ensure privacy with regard to this dataset.

Bibliography

About - UCI Machine Learning Repository. (n.d.). Retrieved February 23, 2024, from https://archive.ics.uci.edu/about

CC BY 4.0 Deed | Attribution 4.0 International | Creative Commons. (n.d.). Retrieved February 1, 2024, from https://creativecommons.org/licenses/by/4.0/deed.en

Cecchi, S., Terenzi, A., Orcioni, S., Riolo, P., Ruschioni, S., & Isidoro, N. (2018, May 14). *A Preliminary Study of Sounds Emitted by Honey Bees in a Beehive*. Audio Engineering Society Convention 144. https://www.aes.org/e-lib/online/browse.cfm?elib=19498

Donation - UCI Machine Learning Repository. (n.d.). Retrieved February 24, 2024, from https://archive.ics.uci.edu/contribute/donation

Functional Requirements for Bibliographic Records (FRBR) – IFLA. (n.d.). Retrieved February 2, 2024, from https://www.ifla.org/references/best-practice-for-national-bibliographic-agencies-in-a-digital-age/resource-description-and-standards/bibliographic-control/functional-requirements-the-frbr-family-of-models/functional-requirements-for-bibliographic-records-frbr/

Hadjur, H., Ammar, D., & Lefèvre, L. (2022). Toward an intelligent and efficient beehive: A survey of precision beekeeping systems and services. *Computers and Electronics in Agriculture*, *192*, 106604. https://doi.org/10.1016/j.compag.2021.106604

Nolasco, I., & Benetos, E. (2021). *To bee or not to bee: Investigating machine learning approaches for beehive sound recognition* (arXiv:1811.06016). arXiv. https://doi.org/10.48550/arXiv.1811.06016

Nolasco, I., & Benetos, E. (2018). *To bee or not to bee: An annotated dataset for beehive sound recognition*. Zenodo. https://doi.org/10.5281/zenodo.1321278

OSBeehives | BuzzBox Hive Health Monitor & Beekeeping App. (n.d.). Retrieved February 2, 2024, from https://www.osbeehives.com/

Repository-Developed Metadata Schemas | DCC. (n.d.). Retrieved February 23, 2024, from https://www.dcc.ac.uk/resources/metadata-standards/repository-developed-metadata-schemas

Re3data.Org. (2014). *UCI Machine Learning Repository*. 588 data sets. https://doi.org/10.17616/R3T91Q

Hart E, Barmby P, LeBauer D, Michonneau F, Mount S, Mulrooney P, Poisot T, Woo KH, Zimmerman N, Hollister JW. 2016. Ten simple rules for digital data storage. *PeerJ Preprints* 4:e1448v2 https://doi.org/10.7287/peerj.preprints.1448v2