FREESOUND DATASETS: A PLATFORM FOR THE CREATION OF OPEN AUDIO DATASETS

Eduardo Fonseca*, **Jordi Pons***, **Xavier Favory***, **Frederic Font, Dmitry Bogdanov, Andres Ferraro, Sergio Oramas, Alastair Porter, and Xavier Serra**
Music Technology Group, Universitat Pompeu Fabra, Barcelona
name.surname@upf.edu

ABSTRACT

Openly available datasets are a key factor in the advancement of data-driven research approaches, including many of the ones used in sound and music computing. In the last few years, quite a number of new audio datasets have been made available but there are still major shortcomings in many of them to have a significant research impact. Among the common shortcomings are the lack of transparency in their creation and the difficulty of making them completely open and sharable. They often do not include clear mechanisms to amend errors and many times they are not large enough for current machine learning needs. This paper introduces Freesound Datasets, an online platform for the collaborative creation of open audio datasets based on principles of transparency, openness, dynamic character, and sustainability. As a proof-of-concept, we present an early snapshot of a large-scale audio dataset built using this platform. It consists of audio samples from Freesound organised in a hierarchy based on the AudioSet Ontology. We believe that building and maintaining datasets following the outlined principles and using open tools and collaborative approaches like the ones presented here will have a significant impact in our research community.

1. INTRODUCTION

Machine learning has a prominent role in data-driven approaches nowadays for many research fields, including sound and music computing. Because of this, having well curated datasets is essential for allowing solid research outcomes. The ImageNet dataset powered most recent advances in computer vision research [5, 27]. This was possible because ImageNet is a large-scale, openly available dataset with a solid ground truth. Despite quite a few datasets being available in the sound and music computing field, there are still major shortcomings in many of them. What follows is a list of some of the most relevant datasets from the sound and music computing community along with some observations. We put a special focus on music and environmental sounds datasets. Table 1 shows some general statistics about them.

• **GTZAN** [39]. Its online availability [1] has enabled easy benchmarking of music genre recognition tasks in the music information retrieval (MIR) field [35]. Despite its popularity, this dataset has been criticized for its small size, its faults and its commonly used data partitions [36]. However, its faults (repetitions, mislabelings, and distortions) were not identified until 2012, ten years after its release. In addition, commonly used data partitions were found to provide an over-optimistic view of the state-of-the-art. Recent work shows that performance is much worse when using a "fault-filtered" GTZAN partition [13].

• **Ballroom** [10]. This music dataset contains beat, tempo and genre annotations [10, 14] and can therefore be used for more than one task. It has also been highly criticized for its small size, its repeated songs (thirteen replicas were found [2]), and the strong relationship between tempo and genre labels (even though the dataset was designed to assess rhythmic descriptors) [10]. Recently, an extension was proposed [20] and 4180 audio clips are now available for 13 unbalanced classes.

• The **Million Song Dataset** [2] was released to provide a large-scale dataset for MIR benchmarking. It contains audio features and metadata for a million contemporary popular music tracks, with a bias towards pop/rock songs. [3] Audio features can be linked to resources useful for several MIR disciplines: lyrics, CD artwork, tags, similarity measures, user data, cover songs or genre labels. This makes it perfect for exploring multimodal approaches. However, the audio files are not available, and the provided audio features were extracted with proprietary software which is neither debuggable nor inspectable [31].

• The **MagnaTagATune** [17] dataset includes music data released under Creative Commons (CC) licenses, which simplifies data sharing, and annotations (tags and similarity) were made by engaging users in playing a game. Since gamification was a research goal in itself, the

*Equally contributing authors.

[1] http://marsyasweb.appspot.com/download/data_sets/

[2] http://media.aau.dk/null_space_pursuits/2014/01/ballroom-dataset.html

[3] http://www.ifs.tuwien.ac.at/mir/msd/MAGD.html

annotation tool is well documented [16]. This makes this dataset more transparent than others in terms of its creation process. As main drawbacks, it is highly unbalanced and its annotations are noisy and inconsistent. [4] To alleviate these issues, researchers typically use top-50 tags [25, 6], or a cleaner and pre-processed version (Magtag5k [21]).

• **AudioSet** [9] is, to date, the largest audio dataset available. It is structured with an ontology containing 632 classes, including music and environmental sound concepts. AudioSet provides a web-interface to navigate and listen to examples from the dataset, which helps giving researchers an overview of its contents. However, even though the dataset annotations have been manually validated, ≈15% of the categories present a quality estimate with a score below 50%. [5]

• **TUT Acoustic Scenes** [23] is a publicly available dataset released for the DCASE 2016 challenge. [6] It is composed of high quality real-world binaural recordings for environmental sound research. A subset of this dataset is also annotated with timestamps and labels for individual sound events. A cross-validation setup is provided together with a baseline. Data collection procedures are explained for possible extension by other parties, and identified errors are listed. The dataset was recorded and annotated by two people, each annotating half of it.

• **UrbanSound8K** [30] includes a taxonomy and two audio collections for urban sounds research. *UrbanSound* includes variable length recordings with timestamps for sound events and salience annotations. *UrbanSound8K* contains labeled slices from these events provided in folds for benchmarking with a baseline. The authors compared several classification models on this dataset and observed that deep models only outperformed shallow models when applying data augmentation techniques [29], suggesting that current machine learning approaches require large and varied datasets.

• **ESC** [24] is an open dataset for environmental sound classification that includes an estimation of human-level performance, a baseline, and code for reproducing author's original classification results. This dataset is composed of two main parts. *ESC-50* contains 2000 annotated clips manually annotated by a single person, while *ESC-US* is a compilation of 250k unlabeled clips. The substantial scale difference between them exemplifies how unscalable annotation procedures can limit the size of datasets.

Based on the observations made in this review, we draw a number of conclusions which could be understood as requirements to consider when creating a dataset: *(i)* small datasets may limit the application of certain machine learning techniques, thus larger datasets are desirable; *(ii)* dataset creation processes must be scalable and sustainable to be able to create large datasets; *(iii)* datasets are sometimes re-annotated and complemented with new data which makes them suitable for new tasks; *(iv)* ways to amend existing datasets and turn them into something dynamic should be established; *(v)* it is important to document the workflows of the dataset creation process and make them transparent; *(vi)* intuitive interfaces for navigating the contents of a dataset are useful for gaining insight; *(vii)* providing data splits facilitates reproducibility and benchmarking; and *(viii)* open licenses allowing for the free distribution of the audio content are desirable for higher research impact.

In this paper we introduce Freesound Datasets, an online platform for the collaborative creation of audio datasets which, based on the requirements above, follows principles of transparency, openness, dynamic character, and sustainability. This paper describes our vision of this platform as a long term project and the first steps that we have carried out as a proof-of-concept. The remainder of this paper is organized as follows. In section 2 we outline the core ideas of our vision and the creation of open audio datasets. Section 3 describes the current state of the platform, which at the time of this writing already allows community contributions through validation of existing annotations. In section 4 we present an early snapshot of a large-scale audio dataset built using this platform and which includes audio samples from Freesound [7] organised in a hierarchy based on the AudioSet Ontology. We end this paper with a summary and future work in section 5.

2. FREESOUND DATASETS VISION

We envision a collaborative process for creating audio datasets [8] built by a community of users that can contribute in different aspects of the dataset creation process. After the observations reported in section 1 and by embracing the ideas described in [22] and [33] for sustainable MIR evaluation and reproducibility of computational methods, we define the following principles that apply to our vision of Freesound Datasets and the creation of datasets with the online platform:

• **Transparency**. It is important that workflows in the dataset creation process are transparent so that dataset users are aware of them. This will allow a better understanding of the dataset itself, its potential and limitations. In this respect, facilitating the exploration of the content through intuitive interfaces is a useful functionality that is often overlooked. Moreover, splits of datasets (e.g., train and test) should be proposed and made publicly available for system benchmarking and reproducibility, so that researchers can carry out experiments whose results are directly comparable.

• **Openness**. It is necessary that datasets are completely open, including audio data and ground truth. Both should be available under open licenses that allow the free distribution and reuse of their content. Further, other relevant data generated during the dataset creation process could be

[4] For example, the following tags are equivalent: *beat/beats* or *female singer/female singing/female vocals/woman singing*.

[5] See `https://research.google.com/audioset/dataset/index.html` for further details on how quality is estimated, accessed 26th April 2017.

[6] `http://www.cs.tut.fi/sgn/arg/dcase2016/`

[7] `https://freesound.org/`

[8] By audio datasets we mean datasets that can include not only audio waveforms but also other audio-related data, e.g., tags or descriptions corresponding to the audio samples.

Dataset (release year)	*#clips*	*clip length*	*dataset duration*	*#classes*	*do authors provide audio? / license type*
GTZAN (2002)	1000	30s	8.33h	10 – balanced	yes / questionable © permission
Ballroom (2006)	698	≈30s	≈5.81h	8 – unbalanced	yes / questionable © permission
Million Song Dataset (2011)	1M	-	-	*external resources*	no
MagnaTagATune (2009)	25,856	≈30s	215.46h	188 – unbalanced	yes / open license
AudioSet (2017)	≈2.1M	10s	≈5833h	527 – (un)balanced*	no
TUT Acoustic Scenes (2016)	1560	30s	13h	15 – balanced	yes / open license
UrbanSound8k (2014)	8732	≤4s	8.75h	10 – balanced	yes / open license
ESC-50 (2015)	2000	5s	2.78h	50 – balanced	yes / open license
FSD early snapshot (2017)	23,519	≤90s	119h	398 – unbalanced	yes / open license

Table 1: Characteristics of the reviewed datasets and presented early snapshot of FSD. *Different partitions are provided.

made available (e.g., annotation procedures and the actual raw annotations). Keeping this information as open as possible aids in the detection of potential issues or biasses in the collection process.

• **Dynamism**. It is desirable that the dataset and the procedures carried out in its collection can be the subject of discussion. We have seen that in some previous works criticism of specific datasets is made [36] and alternative versions or subsets of a dataset are proposed [13]. We envision such criticism and proposals happening in a collaborative online platform where detected faults and issues can be discussed and adequately addressed. This would imprint a dynamic character to datasets which could be versioned and updated with contributions from the community.

• **Sustainability**. For such a vision to stand in the long term, a sustainable approach is required not only in terms of gathering audio content and annotations, but also in terms of maintenance. In the envisioned scenario the community acts as a continuous source of information at different levels. Ideally, the community would be self-sufficient as a source of audio-related content by uploading and sharing open content at large scale. Indeed, some previous works have adopted similar approaches for gathering huge amounts of data based on user-provided content, (e.g., AudioSet, based on YouTube videos [9], or ImageNet, based on Flickr and other search engines [5]). In order to construct corresponding ground truth at a large scale level, it is likely that a substantial part of the annotations needs to be gathered through crowdsourcing. Finally, technical maintenance requirements should be kept as low as possible.

2.1 Objectives

Based on the aforementioned principles, the main goals of the Freesound Datasets platform are: *(i)* to **allow the creation and sharing of open audio datasets** containing audio and/or metadata that the community can leverage, be them of general purpose or tailored to specific research problems. And *(ii)* to **allow room for discussion** around the datasets with the purpose of gaining insight and identifying potential improvements. The discussion will ideally be focused not only on the dataset content but also on the workflow of the data collection process. With respect to the content, datasets are intended to be dynamic in the sense that they can evolve over time at multiple levels. Firstly, detected errors, e.g., mislabelings or distorted sounds, should be amended. Secondly, we also consider the possibility of expanding the datasets when more annotated content is generated. Finally, major modifications of a dataset could also be addressed if firm agreement by the research community exists, e.g., modifications of the taxonomy when applicable. As a result, datasets created in such framework may improve over time in terms of quality (better ground truth annotations), and quantity (larger amounts of data). This concept of time-evolving datasets triggers the need for appropriate dataset versioning, leading to consecutive releases of every dataset. Assigning permanent identifiers, e.g., Digital Object Identifiers (DOIs), to releases can help to enhance their unique identification and proper citation, e.g., using a tool like Zenodo [9] [26].

2.2 Two Key Factors

The success of a platform like the one we envision relies on two key factors: *(i)* the need of substantial and regular sources of open and diverse audio-related information, and *(ii)* the need of a community of users able to enrich datasets by providing annotations. In regards to the first factor, we plan to leverage Freesound, an online collaborative audio sample sharing site that has been supporting diverse research and artistic purposes since 2005 [7]. Freesound has 6.5 million registered users and over 340,000 sounds. More than 3000 new sounds are added every month. [10] The most obvious type of content is audio samples, covering a wide range from music samples to environmental sounds, including human sounds, audio effects, etc. Also, users complement the sounds with metadata, e.g., tags, descriptions and comments. A remarkable characteristic is that quality is prioritized over quantity in terms of sound quality and metadata associated to the sound files. All of the content is released under CC licenses. A number of openly available datasets containing Freesound clips have already been used for research [24, 30, 34], showing that it is already a useful source for the creation of datasets.

Regarding the second factor, we need a community around the datasets to enrich their data. Freesound already has a highly engaged community who contributes to the ecosystem by uploading, rating and discussing sounds. We

[9] `https://zenodo.org/`
[10] Data from 26th April 2017.

believe that part of this community will be interested in contributing to a platform like Freesound Datasets. We also hope that a broader community will form around this initiative, consisting of members of the research community and other sound enthusiasts who share our principles.

3. FREESOUND DATASETS PLATFORM

A working prototype of the Freesound Datasets platform has been already deployed and is available at `https://datasets.freesound.org`. For each hosted dataset, [11] the platform provides a number of tools which are described in the following subsections.

3.1 Annotation Tools

This is the set of tools which allows us to provide annotations about the contents of a dataset (i.e., about its audio resources). These annotations could be labels or any sort of information to be used as ground truth. The current version of the platform only implements one way for users to provide annotations. This is a validation task in which users are presented with a number of audio samples and, for each sample, have to assess the presence of a particular sound category. In future iterations of the platform we intend to support other annotation tasks, e.g., adding labels to audio samples or annotating timestamps of specific acoustic events happening within an audio file.

Several options exist for collecting annotations, such as relying on experts or leveraging annotation effort from volunteers in a controlled environment, e.g., a university campus. However, these options seem neither scalable nor sustainable when collecting large-scale datasets. In this case, regardless of the nature of the annotation task, we expect the bulk of annotations to be provided by the community of platform users in a *crowdsourced* fashion. Many existing datasets have been built via crowdsourcing, ImageNet being an iconic example due to its impact in the image processing field [27]. Different kinds of crowdsourcing approaches have been explored, which mainly vary in the way users are rewarded by their contributions, including volunteering-based approaches, games with a purpose, and *paid-for* crowdsourcing [28]. In the sound and music computing field, a number of initiatives have already explored the use of volunteering-based crowdsourcing approaches [11, 37, 22]. The audio community has also extensively explored the *gamification* approach, where the annotation task is presented as an engaging and entertaining experience. Examples of this include games with the purpose of collecting tag annotations, e.g., TagATune [18], the Listen Game [38] and MajorMiner [19]; or games to collect similarity measurements like Spot the Odd Song Out [41]. Finally, a few *paid-for* crowdsourcing experiences exist in the audio field, e.g., ESC dataset [24] and the VU Sound Corpus [40]—both of which contain annotations for Freesound content—or MoodSwings Turk [32] and SocialFX [42] datasets.

[11] The current version of the platform is hosting an early snapshot of the dataset described in section 4, and is, at present, the only dataset available.

Effective quality control plays an important role in determining the success of any data collection venture, especially for crowdsourcing annotations. A common solution to ensure good quality in the gathered annotations is to rely on redundancy. For instance, correct answers can be identified by applying majority voting, or a quality score for each user or *worker* that contributed annotations can be estimated [12]. To ensure that workers are qualified enough to successfully contribute to an annotation task, a proper training phase is typically designed along with a simple task design with clear guidelines [28, 27]. These aspects are considered in the implementation of the annotation tools of the platform. In particular, in our implemented validation task, the training phase shows descriptions and representative audio samples of the sound category to be assessed and its related categories, in order to help the worker form a judgment before proceeding with the task. The mechanisms of quality control used are inspired by those of CrowdFlower [12] and good-sounds.org [1], and include, among other measures, the periodic usage of verification clips to ensure that submitted responses are reliable.

3.2 Other Tools

As described in the principles presented in section 2, it is important to provide an environment for intuitively exploring the content of a dataset, reporting mistakes, making available alternative versions of the dataset, and discussing any of the elements involved in the creation workflow. To this end, we envision a number of tools for the Freesound Datasets platform which provide such functionalities:

• **Audio exploration**. Dataset content can be explored by browsing the audio samples organized by sound categories and samples can be played while visualizing their waveforms. During this process, it is possible to report faulty audio samples or wrong annotations. Systematically flagged examples can be reallocated in a post-processing stage and, for example, marked for further validation.

• **Data downloading**. A single dataset can be made available for download in different releases which include updated ground truth and contents. Audio samples in their original format are provided, thereby allowing researchers to compute any kind of audio features and to adopt any type of machine learning approach. In addition, audio features [13] pre-computed with the Essentia library [3] are available. Along with the audio content, existing Freesound metadata for audio samples (e.g., user-provided title, tags, textual description, etc.), and collected ground truth data can be retrieved. We plan to link specific releases of datasets with DOIs to facilitate referencing.

• **Discussion tools**. The platform encourages discussion [14] about several aspects of the datasets, including but not limited to: faulty audio samples, wrong annotations, annotation tasks protocol (including aspects such as

[12] `https://www.crowdflower.com/`

[13] A list of pre-computed audio features can be found in `https://freesound.org/docs/api/analysis_docs.html`.

[14] Discussion can be joined at `https://github.com/MTG/freesound-datasets/issues`.

a ground truth target taxonomy), and the platform itself. We decided to host discussions in GitHub, since issues can be created and labeled to organize the discussion in topics, and functionalities to track their evolution are available.

4. EARLY SNAPSHOT OF THE FIRST DATASET

As a proof-of-concept of the use of the Freesound Datasets prototype, we present an early snapshot of a dataset that is being created using the platform. We call this dataset the *FreesoundDataset* (FSD). Similarly to the recently introduced AudioSet [9], FSD aims to be a general-purpose and large-scale audio dataset. Audio samples in FSD are therefore labeled using the same hierarchical ontology of AudioSet, which includes 632 audio classes. In the following subsections we describe how this dataset is being created and discuss its status at the time of this writing.

4.1 Data Gathering and Preprocessing

We started building the FSD by automatically populating it with a number of candidate audio samples from Freesound for each category in the AudioSet Ontology. The selection of candidate audio samples was done based on a process of tag matching in which we manually assigned a number of Freesound tags to each category in AudioSet. Then, each category was automatically populated with all sounds from Freesound that contained the selected tags. Suitable tags were found by considering category descriptions provided in AudioSet and obtaining most frequent Freesound tags that co-occur with the target label. After this first selection of tags, a refinement process was performed in some categories by defining tags to be rejected when needed.

After this initial process, which was carried out by three of the authors, we were able to map more than 300,000 Freesound clips to the AudioSet classes. Because sounds in Freesound can widely vary in length, we decided to filter out all samples longer than 90s, which left us with a total of 268,261 candidate samples in the FSD. Each sample was annotated with an average of 2.62 AudioSet categories.

In order to assess the quality of the selected candidates, we conducted an Internal Quality Assessment (IQA). It consisted of a validation task which was carried out using the Freesound Datasets interface. For each sound category, 12 randomly chosen audio samples were presented and a single rater validated the presence of that category in each sample, with possible responses being "Present", "Not Present", and "Unsure". A quality value for each category could be estimated as the percentage of "Present" responses. The IQA, performed by 11 subjects, was useful to *(i)* determine categories with very low quality, likely due to mapping errors to be improved, and *(ii)* to collect feedback about the Freesound Datasets validation task interface and incorporate improvements for next phases.

Finally, we discarded sound categories for which there were less than 40 assigned audio samples and for which the rate of "Not present" responses from the IQA was larger than 75%. Since this process removed half of the musical genre categories, we decided to omit the rest of them too.[15] This left a total of 398 sound categories, with an average of 1553 candidate audio samples per category.

4.2 Validating Annotations

Having the automatic annotations provided by the tag matching algorithm for each sound category, the goal was then to manually validate these annotations at a significantly larger scale than in the IQA. To this end, we recruited 31 participants (mostly masters and PhD students from our department) and asked them to carry out a validation task very similar to that of the IQA for the selected 398 audio categories. To facilitate the task of validating annotations, participants were asked to validate groups of related categories (e.g., sibling categories). In this way they could get familiarized with specific sections of the ontology and provide more consistent validations [28].

Raters were first instructed to access the online platform and choose one of the available groups of categories. Then, for every category, they had an initial training phase where they acclimatized themselves with the category by looking at its location in the hierarchy and a provided textual description, together with representative sound examples. After that, they were presented with 12 randomly chosen audio samples from that category and asked to rate its presence as: "Present and predominant" (PP), "Present but not predominant" (PNP), "Unsure" (U) or "Not Present" (NP). They were instructed that PP means that the type of sound is either isolated from other types of sounds or with low background noise, whereas PNP implies that the audio clip also contains other salient types of sound and/or strong background noise. We added these two levels of "presentness" as during IQA we observed that, in some audio samples, several sound sources and/or acoustic events co-existed with different salience levels and this made the *Present* option rather ambiguous. A similar approach was used in [30]. Hereafter, "Present" = PP + PNP.

After 12 clips were validated, participants could continue validating annotations of the same sound category and 12 new (non-validated) samples were presented. Audio samples were presented using headphones and in a quiet classroom environment. Along with the playable audio and its waveform, participants were also given links to the specific Freesound page for each audio sample. In case of doubt, they could open the page to take their decision based on the sound metadata provided there. Likewise, they could leave general feedback for every audio category through a text box. After two annotation sessions, we gathered more than 42k validations from our 31 participants.

4.3 Characteristics of the FSD

The early snapshot of FSD consists of a list of audio samples together with labels that determine the sound category/ies they belong to, (out of the 398 previously selected). The main statistics of the current snapshot can be seen in Table 2. The table shows, from left to right, *(i)* the number of candidates/annotations that were generated

[15] This was expected as Freesound does not host music content in the traditional sense of "songs".

	Mapped	Validated*	Present*	PP*	PNP*
Annot.	703,359	42,575	25,365	21,526	3839
Clips	268,261	37,398	23,519	20,206	3679
Hours	986	176	119	92.5	30

Table 2: Main statistics of the current snapshot of FSD: number of annotations, audio clips and hours of audio for several sets of data. Columns with * refer to the 398 selected categories. "Present" is the union of PP and PNP. Despite PP and PNP responses being complementary, an *audio clip* annotated in several categories could receive different subjective ratings for each annotation. This is why "Present" $\leq$ PP + PNP.

using the tag-based mapping, *(ii)* the amount of them that have been validated through the conducted experiments, and *(iii)* the amount that has been validated as "Present", (split into *(iv)* PP and *(v)* PNP). Since the sound categories are non-exclusive, the number of annotations is larger than that of audio samples (as mentioned above, the average number of annotations per sample is 2.62). We consider annotations rated as "Present" (PP + PNP) as the most relevant for building the ground truth of a dataset. Our early snapshot contains 23,519 audio clips (119h of audio) with 25,365 annotations. The lengths of the audio clips in the snapshot are irregularly distributed up to a maximum of 90s (40% of the samples are shorter than 6s and around 78% last less than 30s). The number of validated annotations varies among the 398 sound categories, but all of them have at least 72 validated annotations, as designed in the conducted experiments. 75% of the sound categories contain at least 40 valid audio samples (i.e., with annotations rated as "Present"), whereas 20% of them contain more than 80 valid audio samples.

While audio samples come with labels expressing the presence of a sound category, exact start and end times of event occurrences are not given, (i.e., *weakly* labeled data [15]). However, 7015 clips with annotations rated as PP are also shorter than 4s. [16] Based on these two conditions, we can assume that most of those audio clips are just examples for those acoustic events, and can be considered *strongly* labeled data [15]. Thus, we can estimate that the snapshot is composed mainly of weakly labeled data and a small amount of strongly labeled data, in a rough proportion of 70%/30%. Finally, Figure 1 depicts the number of validated annotations gathered for each of the seven *families* of sounds, according to the first layer of the AudioSet Ontology [9].

4.4 Discussion

The FSD snapshot comprises a wide range of audio samples in terms of content, recording scenarios and sources, which presumably makes it representative of real world situations. The differentiation between PP and PNP allows us to have two different subsets of audio presenting different conditions (see section 4.2). Among the shortcomings, the mapping used to generate candidates for AudioSet categories is not optimal, but it is a starting point that will enable future improvements, e.g., a content-based mapping. Another issue is that, for this early snapshot, annotations were only validated by a single rater. While we believe that annotator agreement is required for defining ground truth based on human-sourced annotations [8], it is also true that a number of datasets do not meet this condition [24,23,39].

[16] 4s is taken as a reference length [30] since it was found to be enough for humans to recognize environmental sounds with 82% accuracy [4].

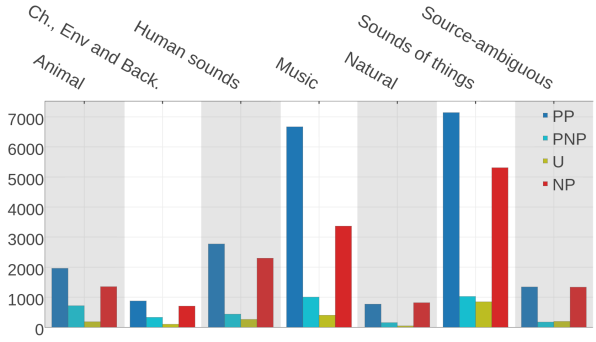

Figure 1: Number of validated annotations (PP, PNP, U, NP) for the different *families* of sounds according to the first layer of the AudioSet Ontology.

Compared to AudioSet, from which FSD takes its ontology, the presented early snapshot is much smaller (Table 1). Currently, AudioSet offers more categories (527) with available content. However, FSD is accompanied by audio waveforms and metadata. Furthermore, FSD includes a mixture of strongly and weakly labeled data whereas in AudioSet only weakly labeled data is provided.

We believe that there exist several applications for FSD within the field of machine listening such as audio event recognition, which enable a variety of specific tasks, e.g., multimedia description, semantically assisted annotation or wildlife monitoring. It also allows a number of approaches like the usage of strongly and weakly labeled data or multimodality, e.g., using audio and metadata for classification. Moreover, future snapshots of the FSD will include improved ground truth data which will further increase its value for research.

5. SUMMARY AND FUTURE WORK

In this paper we have introduced Freesound Datasets, an online platform for the collaborative creation of open audio datasets. We have outlined the core ideas of our vision on the creation of open audio datasets. The current state of the online platform has been described and we have also presented an early snapshot of a large-scale audio dataset built using this platform. This being a long term project, only first steps have been carried out. Next milestones include enhancing the platform functionalities and adding new ones that allow us to crowdsource annotations reliably for new annotation tasks, while promoting discussion around the datasets. After gathering more validated annotations for FSD, we will make the first release including data splits and a baseline for reproducibility and benchmarking. We hope that our platform can serve as an inspiration for creating datasets of completely different nature.

6. ACKNOWLEDGMENTS

This work has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 688382 "AudioCommons", and from the Maria de Maeztu Units of Excellence Programme (MDM-2015-0502). The authors would like to thank Nat Friedman for his support through the AI Grant, Sertan Senturk for his valuable feedback, and the volunteers who participated in the conducted experiments.

7. REFERENCES

[1] Giuseppe Bandiera, Oriol Romani Picas, Hiroshi Tokuda, Wataru Hariya, Koji Oishi, and Xavier Serra. Good-sounds.org: A framework to explore goodness in instrumental sounds. In *International Society for Music Information Retrieval Conference*, pages 414–419, 2016.

[2] Thierry Bertin-Mahieux, Daniel P.W. Ellis, Brian Whitman, and Paul Lamere. The million song dataset. In *Proceedings of the 12th International Conference on Music Information Retrieval (ISMIR 2011)*, 2011.

[3] Dmitry Bogdanov, Nicolas Wack, Emilia Gómez, Sankalp Gulati, Perfecto Herrera, Oscar Mayor, Gerard Roma, Justin Salamon, José R Zapata, Xavier Serra, et al. Essentia: An audio analysis library for music information retrieval. In *ISMIR*, pages 493–498, 2013.

[4] Selina Chu, Shrikanth Narayanan, and C.-C. Jay Kuo. Environmental sound recognition with time-frequency audio features. *Transactions on Audio, Speech, and Language Processing*, 17(6):1142–1158, August 2009.

[5] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proceedings of the Conference on Computer Vision and Pattern Recognition*, pages 248–255. IEEE, 2009.

[6] Sander Dieleman and Benjamin Schrauwen. End-to-end learning for music audio. In *Proceedings of the Acoustics, Speech and Signal Processing International Conference*, pages 6964–6968. IEEE, 2014.

[7] Frederic Font, Gerard Roma, and Xavier Serra. Freesound technical demo. In *Proceedings of the ACM International Conference on Multimedia*, pages 411–412. ACM, 2013.

[8] Peter Foster, Siddharth Sigtia, Sacha Krstulovic, Jon Barker, and Mark D Plumbley. Chime-home: A dataset for sound source recognition in a domestic environment. In *Workshop on Applications of Signal Processing to Audio and Acoustics*, pages 1–5. IEEE, 2015.

[9] Jort F Gemmeke, Daniel PW Ellis, Dylan Freedman, Aren Jansen, Wade Lawrence, R Channing Moore, Manoj Plakal, and Marvin Ritter. Audio set: An ontology and human-labeled dartaset for audio events. In *Proceedings of the Acoustics, Speech and Signal Processing International Conference*, 2017.

[10] Fabien Gouyon, Anssi Klapuri, Simon Dixon, Miguel Alonso, George Tzanetakis, Christian Uhle, and Pedro Cano. An experimental comparison of audio tempo induction algorithms. *Transactions on Audio, Speech, and Language Processing*, 14(5):1832–1844, 2006.

[11] Perfecto Herrera, Òscar Celma, Jordi Massaguer, Pedro Cano, Emilia Gómez, Fabien Gouyon, Markus Koppenberger, David García, J Mahedero, and Nicolas Wack. Mucosa: A music content semantic annotator. In *Proceedings of the International Conference on Music Information Retrieval*, pages 77–83, 2005.

[12] Panagiotis G Ipeirotis, Foster Provost, and Jing Wang. Quality management on amazon mechanical turk. In *Proceedings of the ACM SIGKDD workshop on Human Computation*, pages 64–67. ACM, 2010.

[13] Corey Kereliuk, Bob L Sturm, and Jan Larsen. Deep learning and music adversaries. *Transactions on Multimedia*, 17(11):2059–2071, 2015.

[14] Florian Krebs, Sebastian Böck, and Gerhard Widmer. Rhythmic pattern modeling for beat and downbeat tracking in musical audio. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 227–232, 2013.

[15] Anurag Kumar and Bhiksha Raj. Audio event and scene recognition: A unified approach using strongly and weakly labeled data. *arXiv preprint arXiv:1611.04871*, 2016.

[16] Edith Law and Luis Von Ahn. Input-agreement: a new mechanism for collecting data using human computation games. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1197–1206. ACM, 2009.

[17] Edith Law, Kris West, Michael I Mandel, Mert Bay, and J Stephen Downie. Evaluation of algorithms using games: The case of music tagging. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 387–392, 2009.

[18] Edith LM Law, Luis Von Ahn, Roger B Dannenberg, and Mike Crawford. Tagatune: A game for music and sound annotation. In *Proceedings of the International Symposium on Music Information Retrieval*, volume 3, page 2, 2007.

[19] Michael I Mandel and Daniel PW Ellis. A web-based game for collecting music metadata. *Journal of New Music Research*, 37(2):151–165, 2008.

[20] Ugo Marchand and Geoffroy Peeters. The Extended Ballroom Dataset. Late-Breaking Demo Session of the 17th International Society for Music Information Retrieval Conference. 2016.

[21] Gonçalo Marques, Marcos Aurélio Domingues, Thibault Langlois, and Fabien Gouyon. Three current issues in music autotagging. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 795–800, 2011.

[22] Brian McFee, Eric Humphrey, and Julián Urbano. A plan for sustainable mir evaluation. In *Proceedings of the International Society for Music Information Retrieval Conference*, pages 285–291, 2016.

[23] Annamaria Mesaros, Toni Heittola, and Tuomas Virtanen. TUT database for acoustic scene classification and sound event detection. In *Proceedings of the European Signal Processing Conference*, pages 1128–1132, 2016.

[24] Karol J Piczak. Esc: Dataset for environmental sound classification. In *Proceedings of the ACM International Conference on Multimedia*, pages 1015–1018. ACM, 2015.

[25] Jordi Pons, Olga Slizovskaia, Rong Gong, Emilia Gómez, and Xavier Serra. Timbre analysis of music audio signals with convolutional neural networks. *arXiv preprint arXiv:1703.06697*, 2017.

[26] Stefan Pröll and Andreas Rauber. Enabling reproducibility for small and large scale research data sets. *D-Lib Magazine*, 23(1/2), 2017.

[27] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[28] Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *Proceedings of the International Conference on Language Resources and Evaluation*, pages 859–866, 2014.

[29] Justin Salamon and Juan Pablo Bello. Deep convolutional neural networks and data augmentation for environmental sound classification. *IEEE Signal Processing Letters*, 24(3):279–283, 2017.

[30] Justin Salamon, Christopher Jacoby, and Juan Pablo Bello. A dataset and taxonomy for urban sound research. In *Proceedings of the ACM International Conference on Multimedia*, pages 1041–1044. ACM, 2014.

[31] Alexander Schindler and Andreas Rauber. Capturing the temporal domain in EchoNest features for improved classification effectiveness. In *International Workshop on Adaptive Multimedia Retrieval*, pages 214–227. Springer, 2012.

[32] Jacquelin A Speck, Erik M Schmidt, Brandon G Morton, and Youngmoo E Kim. A comparative study of collaborative vs. traditional musical mood annotation. In *International Society for Music Information Retrieval Conference*, pages 549–554, 2011.

[33] Victoria Stodden, Marcia McNutt, David H Bailey, Ewa Deelman, Yolanda Gil, Brooks Hanson, Michael A Heroux, John PA Ioannidis, and Michela Taufer. Enhancing reproducibility for computational methods. *Science*, 354(6317):1240–1241, 2016.

[34] Dan Stowell and Mark D Plumbley. An open dataset for research on audio field recording archives: freefield1010. *arXiv preprint: 1309.5275*, 2013.

[35] Bob L Sturm. A survey of evaluation in music genre recognition. In *International Workshop on Adaptive Multimedia Retrieval*, pages 29–66. Springer International Publishing, 2012.

[36] Bob L Sturm. The gtzan dataset: Its contents, its faults, their effects on evaluation, and its future use. *arXiv preprint: 1306.1461*, 2013.

[37] Nikolaos Tsipas, Charalampos A Dimoulas, George M Kalliris, and George Papanikolaou. Collaborative annotation platform for audio semantics. In *Audio Engineering Society Convention 134*. Audio Engineering Society, 2013.

[38] Douglas Turnbull, Ruoran Liu, Luke Barrington, and Gert RG Lanckriet. A game-based approach for collecting semantic annotations of music. In *Proceedings of the International Conference on Music Information Retrieval*, volume 7, pages 535–538, 2007.

[39] George Tzanetakis and Perry Cook. Musical genre classification of audio signals. *Transactions on Speech and Audio Processing*, 10(5):293–302, 2002.

[40] Emiel van Miltenburg, Benjamin Timmermans, and Lora Aroyo. The VU sound corpus: adding more fine-grained annotations to the freesound database. In *Proceedings of the International Conference on Language Resources and Evaluation*, 2016.

[41] Daniel Wolff. *Spot the Odd Song Out: Similarity Model Adaptation and Analysis using Relative Human Ratings*. PhD thesis, City University London, 2014.

[42] Taylor Zheng, Prem Seetharaman, and Bryan Pardo. Socialfx: Studying a crowdsourced folksonomy of audio effects terms. In *Proceedings of the 2016 ACM on Multimedia Conference*, pages 182–186. ACM, 2016.