Deliverable D5.5

Evaluation report on the tool for manual annotation of non-musical content

Grant agreement nr	688382
Project full title	Audio Commons: An Ecosystem for Creative Reuse of Audio Content
Project acronym	AudioCommons
Project duration	36 Months (February 2016 - January 2019)
Work package	WP5
Due date	31st July (M30)
Submission date	31st July (M30)
Report availability	Public (X), Confidential ( )
Deliverable type	Report (X), Demonstrator (), Other ()
Task leader	MTG-UPF
Authors	Xavier Favory, Eduardo Fonseca, Frederic Font
Document status	Draft (), Final (X)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

Table of contents

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N° 688382

Executive Summary

This deliverable presents the Audio Commons Manual Annotator, a tool for the manual annotation of audio content. It is a web-based interface that intelligently guides users on the annotation process of a large variety of sound concepts. This tool is being integrated into Freesound Datasets, a platform for the creation of open audio datasets.

One of the challenges in making use of Creative Commons audio content comes from the fact that it is provided by various sources and authors with different backgrounds and levels of expertise. Therefore, the content is often unstructured and not properly annotated, which hinders its efficient retrieval. Moreover, there is a scarcity of tools and agreed methods to aid users in the task of annotating audio content through established common procedures. Intelligently guiding users on the annotation process would allow a reliable, uniform and complete description of the content which will therefore facilitate its sharing.

In the first section of this document, we motivate the need of novel tools for the manual annotation of audio content. More specifically, we focus on annotating content with a large set of predefined concepts. Based on the preliminar evaluation carried out on the prototype presented in the deliverable D5.4 Release of tool for the manual annotation of non-musical content, we establish the need to provide a simpler and more focused tool that includes a number of improvements. We call this tool the Audio Commons Manual Annotator.

In the second section, we introduce the Audio Commons Manual Annotator, a web-based tool for the manual annotation of audio content - which guides the user in the process of annotating audio samples with a wide range of sound categories. We then present its evaluation carried out with 4 users, for which we applied a mixed methods approach combining human-computer interaction (HCI) metrics with behavioral and qualitative data analysis. We propose a topic-oriented discussion about the challenges arisen and possible solutions when annotating audio content in a post-process scenario such as the Freesound Datasets platform. Finally, we end this report with a summary of the work done and sketch the next steps to be carried out for the integration of the tool in a crowd-sourcing scenario.

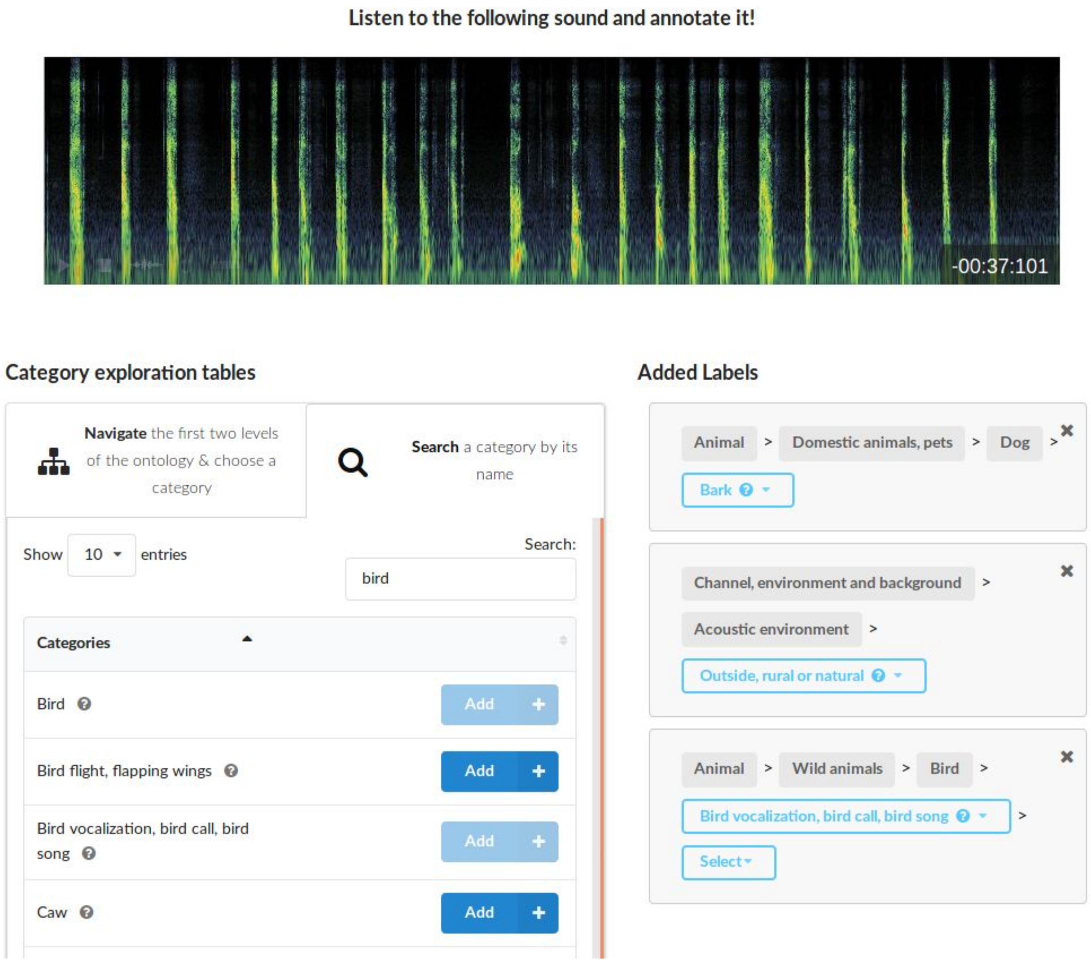
This deliverable is complemented by Task 4.9 which focus on evaluating another tool for the manual refinement of pre-assigned labels.

1 Motivation

In a previous deliverable we presented a prototype for the manual annotation of non-musical content (D5.4 Release of tool for the manual annotation of non-musical content). There, we proposed a web-based interface which guides the user in the process of annotating audio samples. The tool was specifically designed to allow the annotation of audio content with a large vocabulary of sound concepts, which is essential in the annotation of non-musical content.

In particular, we used the recently released AudioSet Ontology as use case - a hierarchically structured collection of 632 categories of everyday sounds [Gemmeke17]. This ontology has been well received by the research community and we believe that it is becoming the de facto standard for various research tasks. Among the main challenges that this type of annotation entails we find an extremely large vocabulary, the heterogeneity of the aspects to annotate, and the variable degree of specificity. This means that the understanding of every category will depend on cultural and/or academic factors (see D5.4 for details).

To address these issues, we proposed in D5.4 an interface that combines two different approaches, as can be seen in Figure 1: (i) the user first explores the taxonomy of concepts and selects labels; (ii) the user is presented with the added labels and then she can remove, refine or specify them.



Listen to the following sound and annotate it!

Navigate the first two levels of the ontology & choose a category

Search a category by its name

Show 10 entries

Search: bird

-00:37:101
Category exploration tables
Added Labels
Categories
Bird
Add +
Bird flight, flapping wings
Add +
Bird vocalization, bird call, bird song
Add +
Caw
Add +
Animal > Domestic animals, pets > Dog >
Bark
Channel, environment and background >
Acoustic environment >
Outside, rural or natural
Animal > Wild animals > Bird >
Bird vocalization, bird call, bird song >
Select

Figure 1. Screenshot of the annotator interface of D5.4 implemented in the Freesound Datasets platform.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

A preliminary evaluation of the tool proposed in D5.4 was carried out with 3 human subjects. After informal discussions with the participants we realized that the tool was trying to account for two different problems at the same time: i) the lack of labels, and ii) the lack of specificity in the labels. Thus, the exploration tables (Figure 1 left) attempt to facilitate the *generation* of new labels, while the *refinement* stage (Figure 1 right) allows to further specify the labels. Both problems lead to unstructured and not properly annotated Creative Commons licensed audio content and hence these are important issues that hinder its potential for sharing and retrieval.

Interestingly, our informal evaluation revealed that tackling both problems at the same time made the tool significantly complex to use, thus requiring a substantial effort on the user side. It must be noted that the ontology under consideration provides a hierarchy of over 600 concepts distributed along six levels of depth, many of which are totally unfamiliar to the annotator. However, apart from the noted complexity, the evaluation of the tool also indicated its potential. All the interface functionalities were found useful by the users and they indeed utilized them and took advantage of all of them.

In light of the above, we decided to split the tool presented in D5.4 into two independent tools, each of them focused on one the aforementioned problems: the *generation* of labels and the *refinement* of labels. By doing this we intend to: i) simplify each tool by narrowing down the purpose of the task, thereby easing its usage; ii) improve each tool by adding functionalities specific for the task at hand.

This document describes and evaluates the generation tool, that is, a tool that helps the user in the process of generating new labels for the audio content. This tool can come in handy during the process of publishing audio content in the Audio Commons Ecosystem (e.g., when content creators upload content to Freesound). As seen in D5.4, annotating non-musical content can require a plethora of very diverse sound aspects. Organizing these aspects with a well-established, rich-enough taxonomy, such as that of AudioSet, will promote uniformity and consistency in the labeling. In turn, this will ease the retrieval and reusability of the audio content. Therefore, one of the main challenges for the tool is to facilitate the exploration and assimilation of concepts in a large-scale ontology like AudioSet, in order to lower the barrier for using the tool. Finally, it is worth mentioning that the tool presented in this document will be deployed in the Freesound Datasets platform [Fonseca17][1], a platform for the collaborative creation of open audio datasets labeled by humans. This resource will be instrumental in the annotation of audio content for ground truth generation.

The refinement tool is described and evaluated in D4.9 and its goal is to aid the user in the task of further specifying a set of already existing labels.

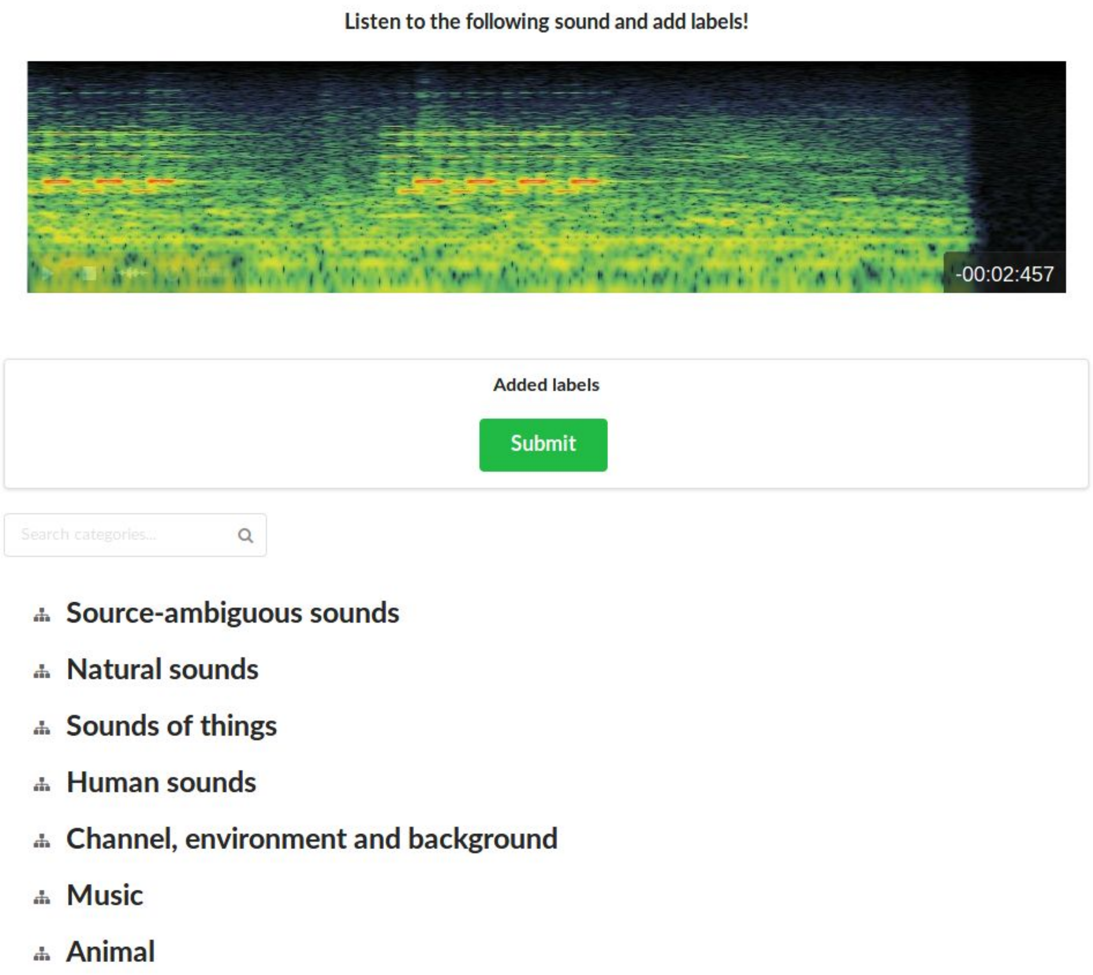
[1] https://datasets.freesound.org/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

2 The AC Manual Annotator

The informal evaluation of the prototype presented in D5.4 revealed that users like to combine the usage of both exploration tables: one for navigation of the ontology and the other to focus on the search of terms within the list of sound concepts. Hence, in the new and improved generation task, we propose a better exploration table which would combine the navigate and search functionalities. More specifically, a tree-like display is designed where the current location in the hierarchy is represented at all times. This gives a more general and clear view of the category context, which is a key factor when deciding the appropriate labels. It also useful for getting acclimatized with the ontology.

Annotation of audio content can be useful for several use cases, e.g., when a provider publishes content in the Audio Common Ecosystem, or in a post-processing stage, where users collaboratively annotate the content, like for instance in the Freesound Datasets platform. The Audio Commons Manual Annotator allows to focus on a single sound resource at a time, which is accessible from a player displaying the spectrogram of the sound in order to facilitate the localization and understanding of sound events in the clip (Figure 2). Labels can be assigned to the audio clip from a button accessible from a category description section, shown in Figure 4. The taxonomy table allows users to open parts of the taxonomy in order to visualize children categories. For each category, textual descriptions are shown, along with sound examples when available. A text-based search, shown in Figure 3 allows to locate categories in the taxonomy table.



Listen to the following sound and add labels!

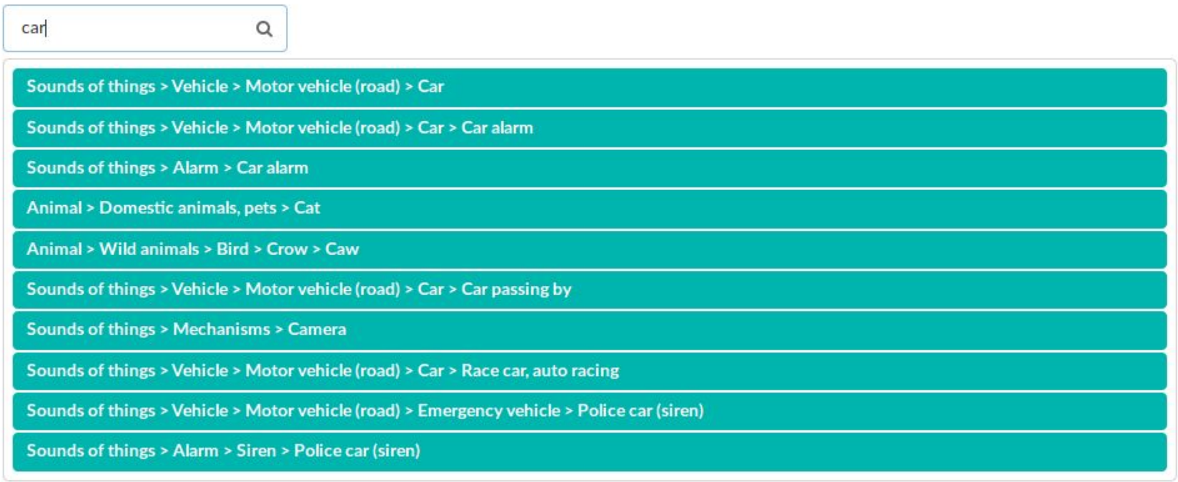
- Source-ambiguous sounds
- Natural sounds
- Sounds of things
- Human sounds
- Channel, environment and background
- Music
- Animal

Figure 2. Screenshot of the Audio Commons Manual Annotator.

A typical workflow would consist in:

1. Listen to the sound sample (Figure 2, top)
2. Use the text-based search to locate categories in the taxonomy table (Figure 3)
3. Explore the taxonomy table to understand well the located category, and perhaps find other more relevant categories (Figure 4)

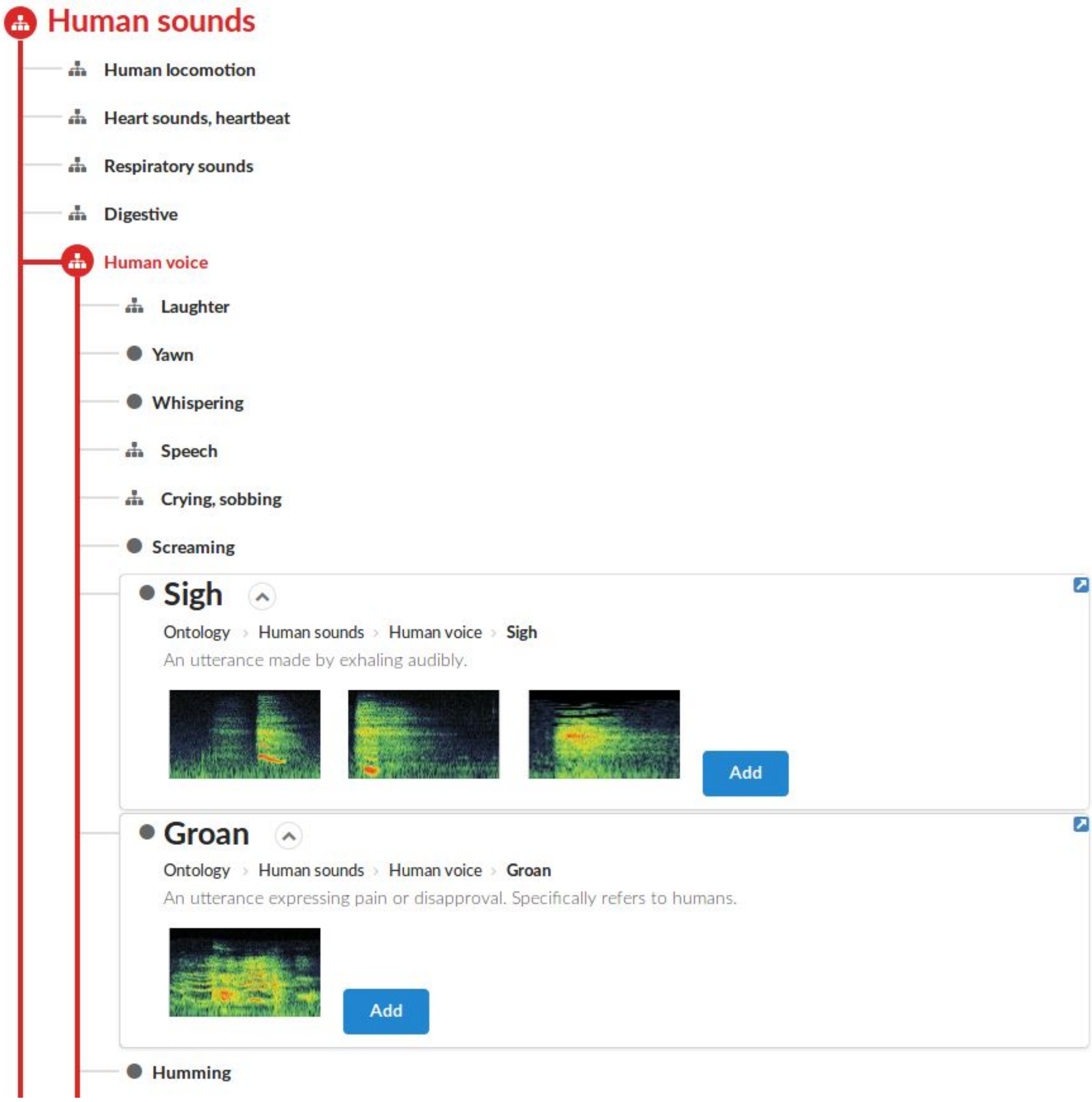
This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382



car

- Sounds of things > Vehicle > Motor vehicle (road) > Car
- Sounds of things > Vehicle > Motor vehicle (road) > Car > Car alarm
- Sounds of things > Alarm > Car alarm
- Animal > Domestic animals, pets > Cat
- Animal > Wild animals > Bird > Crow > Caw
- Sounds of things > Vehicle > Motor vehicle (road) > Car > Car passing by
- Sounds of things > Mechanisms > Camera
- Sounds of things > Vehicle > Motor vehicle (road) > Car > Race car, auto racing
- Sounds of things > Vehicle > Motor vehicle (road) > Emergency vehicle > Police car (siren)
- Sounds of things > Alarm > Siren > Police car (siren)

Figure 3. Screenshot of the Audio Commons Manual Annotator text-based search input showing the results obtained with the query “car”.



Human sounds

- Human locomotion
- Heart sounds, heartbeat
- Respiratory sounds
- Digestive
- Human voice
  - Laughter
  - Yawn
  - Whispering
  - Speech
  - Crying, sobbing
  - Screaming
  - Sigh
    Ontology › Human sounds › Human voice › Sigh
    An utterance made by exhaling audibly.
    Add
  - Groan
    Ontology › Human sounds › Human voice › Groan
    An utterance expressing pain or disapproval. Specifically refers to humans.
    Add
  - Humming

Figure 4. Screenshot of the Audio Commons Manual Annotator taxonomy table, showing the descriptions and examples of “Sigh” and “Groan”, together with their hierarchy location.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N° 688382

3 Experiment

The main characteristic of collaborative multimedia collection is that the content is provided by different people from different backgrounds and expertises. This tends to produce content that is not uniformly annotated, and does not allow its efficient retrieval. To face this issue, crowdsourcing has emerged as a powerful tool for making the process of annotating large sound collection scalable. In these contexts, there is a need for proposing new manual interfaces to properly annotate audio content, with labels that are comparable and of same nature. In this experiment, we advance our user-driver design process of proposing new annotation tools for annotating audio content from a large variety of types. We take advantage of the AudioSet Ontology which provides a hierarchical taxonomy of very broad acoustic categories. We use the Audio Commons Manual Annotator as a technology probe to observe its use in a real context, to evaluate its functionalities and to inspire new ideas [Hutchinson03]. One of our goals is to propose a method that will guide people in providing annotations that are as consistent as possible.

3.1 Methodology

3.1.1 Task

We selected some sounds from the Freesound Datasets platform featuring one or more of the following aspects: (i) containing multiple sources, (ii) presenting background noise or (iii) hard to recognize. This process resulted in a list of 9 sounds that the user had to fully annotate with the Audio Commons Manual Annotator. At the end of the task, they were provided a questionnaire, followed by a semi-structured interview.

3.1.2 Context, participants and procedure

We gathered four participants with different level of expertise. We will use A, B, C and D letters for referring to them in this document. A is very familiar with Freesound content and the challenges around its accurate annotation. B and C have a bit of experience in using other annotation tools for the annotation of audio content. D is rather not very familiar with this sort of work. All the participants were non native english speakers, but declared to have a excellent level.

Some guidelines were shown to them, together with explanations given by the examinator. First the context of the study was explained: evaluate novel interfaces for the manual annotation of audio content with large vocabulary of sound related concepts. They were instructed, for every audio clip, to select all the labels that apply using the taxonomy table. The taxonomy was presented as a hierarchical structure containing over 600 audio categories. Upper levels in the hierarchy contain broader sound concepts while lower levels are formed by more specific categories. These categories mostly include concepts related to: (i) Sound events, source or production mechanisms (e.g., 'Bark', 'Tearing', …); (ii) Categories describing aspects of sound (e.g., 'Reverberation', 'Boing', ...). They were asked to specify the labels as much as possible by going deeper in the hierarchy, and they were advised to use the text search input for locating the categories in the hierarchy.

While they were performing the task, the participants were asked to think out loud, and share their comments or doubts. The examiner was present during all the experiment, making sure no major issue was avoiding the participants to perform the task, to support them in case of doubts., and to transcribe relevant participants actions and comments.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

3.1.3 Survey

The survey was divided in 2 parts. It first included usability related questions (SUS usability scale) [Brooke96], and then overall feedback on engagement and learning. The SUS questionnaire investigates dimensions related to interest, complexity, ease of use and simplicity/difficulty, integration and consistency. The overall feedback asses the English language level of the participant, the levels of engagement, learning, novelty and quality of category retrieval.

3.1.4 Interview

We conducted semi-structured interviews with the 4 participants after they completed the task and the survey. We used open-ended questions and specific questions related to observed behaviours during the performance of the task. We used thematic analysis to identify emerging themes from participants' answers. These are the questions that were asked to participants, from the which some interesting discussions emerged:

- How intuitive was the interface?
- What about the different features?
  - Search
  - Taxonomy tree
- How difficult was the task?
- Did you have any doubt? How did you react? Could you solve it?
- Would you find it useful to see the sounds metadata from Freesound?
- Would it help to have consecutive sounds with similar labels (one topic)? Like choosing a family?

3.2 Results

3.2.1 Survey

Usability

No significant differences were found that correlate with the level of expertise of the four participants. Figure 5 illustrates the results obtained to the usability questionnaires (SUS). Participants strongly agreed that the interface was easy to use and did not appear unnecessarily complex. They also tended to agree to items related to confidence in using the system, the satisfactory integration of the functions and the quick learning of its use. As a result, the Audio Commons Manual Annotator obtained on overall a high usability score according to the SUS metric (M=82.5, SD=10.75).

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

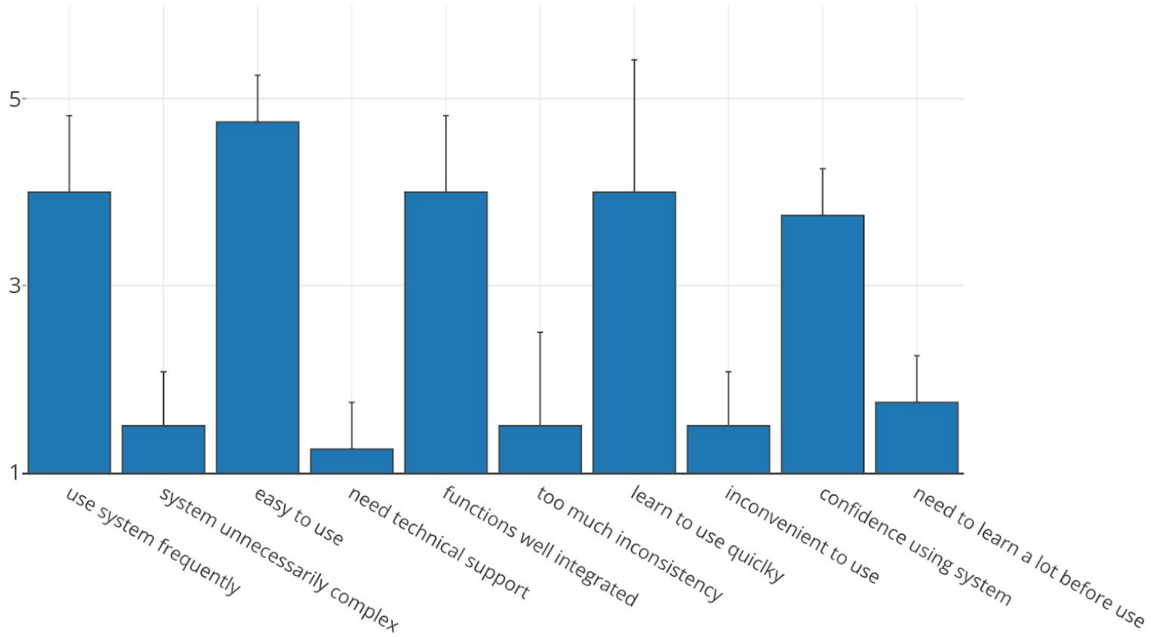


Figure 5. Mean and standard error for the usability questionnaire items (SUS). Value of 1 corresponds to Strongly disagree, 3 to neutral, and 5 to strongly agree.

Figure 6 shows the individual usability score for the four participants. The participant C gave an a usability score that was significantly lower than the others. During his performance of the task, it was observed that he often did not identify the existing sound sources, and spent quite a long time searching for categories in the taxonomy.

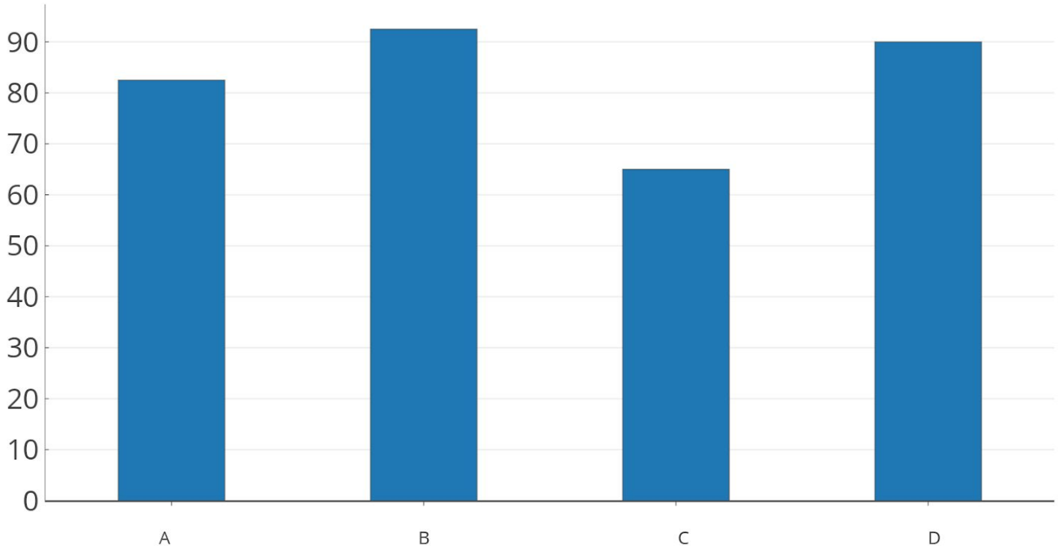


Figure 6. Individual overall usability scores for the four participants

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

Engagement

Further analysis was conducted to analyse responses of participants about learning and satisfaction related to category retrieval. Figure 7 shows the mean and standard error of the different questions. Participants did not totally agree on the amount of knowledge they learnt by performing the task with the proposed interface. However, they were all quite satisfied by the retrieval performance when using the tool.

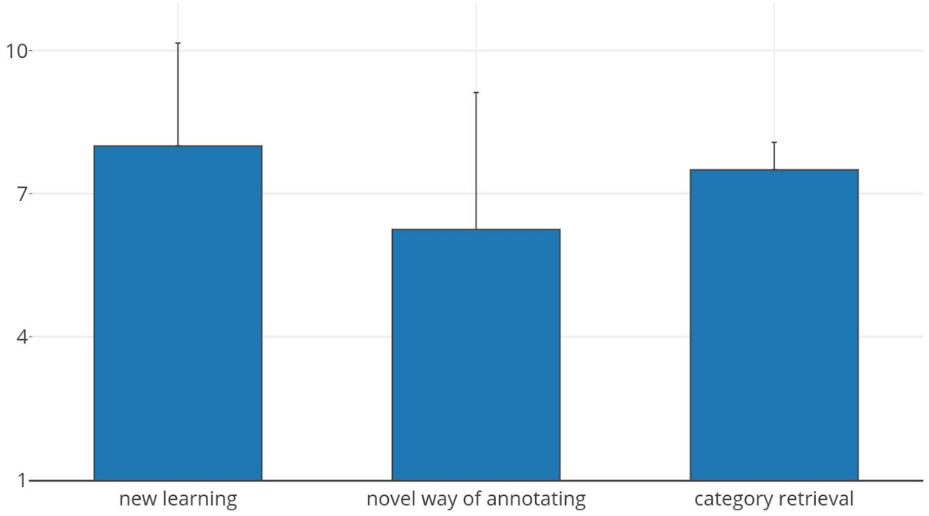


Figure 7. Mean and standard error for the overall feedback questionnaire.

3.2.2 Produced labels

As one of the main idea is to investigate how different annotators from different level of expertise annotate sounds, we propose some statistics that allows us to quantify it. We observe in Table 1 that all the annotators produced a similar number of labels for each sound.

Annotator	Total number of labels produced (9 sounds)	Time spent annotating
A	26	30 minutes
B	21	25 minutes
C	26	25 minutes
D	23	30 minutes

Table 1. Number of total labels produced for the 9 sounds.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

However, when comparing the provided labels, only few are common among participants. Table 2 shows the average number of provided labels that are common between pairs of annotators, e.g. common to users A and B, to users B and C, and the rest of pairs among the four users A, B, C and D. This suggests that several annotators are needed when annotating content with such a large taxonomy of concepts.

Average of pair annotator common labels	Total number of different labels produced
6.5	70

Table 2. Average pair annotator commons labels produced and the total of different labels produced

3.2.3 Interviews and transcriptions

In this section we provide some of the feedback gathered from the interviews and transcriptions of the participants' tasks that will be discussed in the next section.

- Some sounds are hard to recognize. Possibility to help by recommending labels, or showing some metadata (specific to the post annotation of Freesound content)
- When sounds are hard to recognize, people tend to go for "ambiguous"/"abstract" categories that do not convey the identity of the source, but rather characteristics of the sound.
- Some participants use categories that sound similar to the one they heard, but that are not the appropriate one.
- The task may require a bit of training, mostly due to the ontology complexity. Starting with familiar sounds could help to familiarize the annotator with the ontology. Having consecutive similar sounds would also help.
- Search is useful when the sound category is easily recognizable. The exploration table helps when participants are unsure about what to look for.
- People tend to focus a lot on the provided category examples.
- Spectrogram is useful to locate and re-listen carefully parts of the sound. It also allows users to find sounds that could be missed without seeing frequencies' energy in the spectrogram.
- Sound to annotate could be more accessible to allow easy comparison with category sound examples
- Some type of labels were not used at all by some participants. Maybe because they did not know about their existence, Some specific guidance questions, or illustrating examples could help users.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

4 Discussion

4.1. Difficult nature of the task

Hard to recognize sound identity

In the context of post-process annotation of audio content the annotator is typically not the publisher of the content. Hence the annotator usually does not know how the recording conditions were, or what sources were captured. Listening to the sound does not necessarily lead to the identification of the source. We observed that people tend to react in different manners to this difficulty.

First, most of the annotators try to make a huge effort to recognize the source, by listening several time to the audio sample. They make great use of the spectrogram for locating parts that they struggle with. Then, they start exploring the taxonomy table and very often make use of the sound examples provided to compare the sample they are annotating with the different categories in the taxonomy. Even with familiar sounds that they experience in their everyday life, they use examples to feel more confident when assigning the categories.

When the source is not identified, people tend to use abstract categories that the taxonomy provides. For instance, the onomatopoeic categories were often used in this case. People assigned labels that convey characteristics of the source rather than its identity. We also observed an annotator who used other well defined categories that sound like the source in the content being annotated.

Complexity of the categories

The taxonomy used in the task is presented as a hierarchical structure containing over 600 audio categories. Upper levels in the hierarchy contain broader sound concepts while lower levels are formed by more specific categories. The nature of the categories included varies to a high extent, as detailed in D5.4. They can refer to different aspects of the sound. This leads people to focus on different characteristics of the sound to provide a comprehensive annotation of the content.

Time consuming

The task of generating label for an audio sample is very time consuming. They spent around 25-30 minutes for annotating 9 sounds. Thus, the interface should require a minimum set of actions when combining exploration of the taxonomy, searching, listening and addition of the labels. However, screen size is a major issue, and it is important to find the right balance between compactness and clarity.

4.2 Useful features

Choosing the sound examples of a category is of high importance since they are often used for comparing and making the final decision on adding labels to an audio content. Selecting these examples does not consist in just selecting any examples. The examples provided should cover all the variability of a sound class. For a category, there should be the minimum amount of examples to avoid wasting time listening to them for too long, and all of them should show a different instance of a sound category.

It is essential to provide ways for efficiently browsing and exploring such an extensive set of audio categories. Text-based search provides a way for people to find categories with their own words. This is particularly efficient when the annotator recognizes the sources and want to quickly add the corresponding audio category to the content. We used text from the category names and descriptions and perform some trigram based queries[2]. Improving the retrieval system would be very beneficial. For instance, one annotator proposed to add some of the children of the retrieved categories to the results. This option was tried when developing the search engine but was discarded because it tended to add a lot of results which could make the localisation of the relevant categories harder. Moreover, we could also use external lexical resources such as WordNet[3] or Wikipedia[4] to improve the recall of the system, by using respectively synonyms terms or page content terms.

However, text-based search can fail when the annotator does not know the vocabulary. He can then rely on the hierarchy structure of the categories. Tree visualization is a direct representation of it, and can help by allowing to iteratively define more precise concepts by starting with upper level of the taxonomy. Tables are a natural way for browsing collections of items. The taxonomy table we provided aims at combining tree and table structures in order to allow efficient and fast exploration of the categories. Moreover, locating similar categories close from each other helps to refine and validate the choice of a category. Some categories are almost identical and differ only in small details. Also, it is important to assign categories that are as precise as possible, by going deep in the hierarchy, as these categories are more specific and give more annotation value.

Since the set of categories we use can refer to various aspects of the sound, it would be very helpful to provide some guidance to the annotators. Specific questions could be asked in order to make them focus on the different characteristics of the sound that the taxonomy allows to describe. This can take advantage of the work done in the deliverable D5.4 Release of tool for the manual annotation of non-musical content, which state the different category types that are present in the AudioSet Ontology. Among them, sound events and production mechanisms are the most predominant category types, but the taxonomy also include categories describing properties of sounds or acoustic scenes and contexts.

One of the difficulties in the context of post-processing annotation of audio content is the fact that the annotator is not always able to recognize the sound sources. However, in our case, the content comes from Freesound, and is often accompanied by rich metadata including a title, some tags and a description. This information can help annotators to understand the context and provide more accurate annotations. Making this metadata accessible to annotators was something discussed during the interviews. Participants were all very interested by the idea. However, they argued that this should be not given at start, and should rather be an aid that should be requested after having spent a certain effort on analyzing the audio content. Providing directly the metadata would correspond more to a transcription task, were annotators could only focus on the metadata, and forget some of the sound aspects that the metadata fail to convey.

Some other problems are more related to the interface, and its design. Many annotators often compare category examples with the audio sample being annotated. It would be very handy to make the sound player more accessible. Annotators were wasting time scrolling up and down the page. Moreover, some annotators proposed to make the submit button and the selected tags always visible, so that they don't get lost and do not forget what they already added.

[2] This is a feature that Postgres (our database backend) implements: https://www.postgresql.org/docs/9.6/static/pgtrgm.html
[3] https://wordnet.princeton.edu/
[4] https://www.wikipedia.org/

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 688382

5 Conclusions and Future Work

In this deliverable we presented the Audio Commons Manual Annotator, which is used for the annotation of audio samples with labels from the AudioSet hierarchical Ontology. We evaluated it using a mixed methods approach combining HCI metrics with behavioral and qualitative data analysis. The results show that the four participants of the study found the system easy to use. The tool seemed to facilitate browsing and using large taxonomies of concepts for annotating audio content. The use of a fixed vocabulary for the annotation of the content allows to gather more consistent annotations across different annotators. However, it was found that the annotators tend to use different categories to describe the same audio content. It is reasonable to think that this comes from the taxonomy complexity and the semantic overlap that the categories present. Since we evaluated it in a post-process scenario, we see advantages in including this as a new annotation tool in Freesound Datasets, our platform for the collaborative creation of open audio datasets. This new tool can be combined with the validation task already present in the platform, and the tool presented and evaluated in D4.9.

Future work include some improvements on the design, such as making the sound sample more accessible to listen to, or making the added label section more reachable, so that users can annotate faster while getting less confused in the process of selecting labels. Moreover, some basic instructions about how to use the tool must be designed. In addition, specific indications to make users focus on specific sound aspects would help them to produce more exhaustive annotations. Another aspect worth to focus on is the search engine. The retrieval method could be more fitted to the task context by customizing the postgres-based queries. Finally, since this tool will be integrated in the Freesound Datasets platform, quality control mechanisms must be designed to make it suitable for crowd-sourcing.

4 References

[Fonseca17] Fonseca, Eduardo et al. (2017). “Freesound Datasets: A platform for the creation of open audio datasets”. In: Proceedings of the International Society for Music Information Retrieval Conference.

[Gemmeke17] Gemmeke, Jort F et al. (2017). “Audio Set: An ontology and human-labeled dataset for audio events”. In: Proceedings of the Acoustics, Speech and Signal Processing International Conference.

[Brooke96] Brooke, J. (1996). SUS-A quick and dirty usability scale. Usability evaluation in industry, 189(194), 4-7.

[Hutchinson03] Hutchinson, H., Mackay, W., Westerlund, B., Bederson, B. B., Druin, A., Plaisant, C., ... & Roussel, N. (2003, April). Technology probes: inspiring design for and with families. In Proceedings of the SIGCHI conference on Human factors in computing systems (pp. 17-24). ACM.

This project has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement N° 688382