Deliverable D5.4

Release of tool for the 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	(M27)
Submission date	(M27)
Report availability	Public (X), Confidential ( )
Deliverable type	Report ( ), Demonstrator (), Other (X)
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 is complementing the deliverable D4.6 Release of tool for the manual annotation of musical content. It presents the first prototype of a tool for the manual annotation of non-musical content present in the Audio Commons Ecosystem. This tool is a web-based interface that intelligently guides users on the annotation process of a large variety of sound concepts. The base implementation is available under an Apache 2.0 license at this location. Moreover, this tool is being integrated into the Freesound Datasets platform, 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 help users with 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 summarize the particularities of annotating non-musical audio content. We briefly describe some of the main existing taxonomies and ontologies that have been proposed for its organization, and we present the AudioSet Ontology, which can be integrated in the Audio Commons Ontology. We conclude the first section by listing the challenges encountered when annotating non-musical content, specifically when using a large vocabulary. In the second section, we present the current Audio Commons Annotator prototype - a web based tool for the manual annotation of non-musical content - which intelligently guides the user in the process of annotating audio samples with a wide range of sound categories. We end this report with a summary of the work done and sketch the next steps to be carried out for the progressive improvement of the version presented here.

This deliverable will be complemented by Task 4.9 and Task 5.5, where both of the developed manual annotation tools will be evaluated with real users.

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

1 Annotation of non-musical content

In a previous deliverable we presented a prototype for the manual annotation of musical content (D4.6 Release of tool for the manual annotation of musical content). There, we proposed a list of nine content type categories for music samples and pieces to be considered in the manual annotation process (e.g., note, chord, etc), and we highlighted the musical properties that are more relevant for them. For example, for a musical note, there is a clearly defined number of tones that can be generated from a closed set of instruments, i.e., sound sources.

The annotation of non-musical content, however, presents an important difference with respect to the musical case: the great heterogeneity of the content types that can be considered. Non-musical content could be virtually anything not generated by a musical instrument, i.e, a loosely defined set of acoustic material. This material could range from environmental sounds produced by natural phenomena (e.g, rain) or animals (e.g., bird), to sounds made by things (shattering glass) or machines in urban or domestic contexts (jackhammer or kettle whistle). Another interrogant is which sound attributes are the most relevant for this acoustic content. Again, this is highly varying as it depends on the content type. For example, the sound attributes that are appropriate for bird sounds may differ to a high extent from those that apply to jackhammer sounds.

1.1 Organization of non-musical content

Due to this lack of definition, a number of taxonomies and ontologies have been proposed in the sound literature with the aim of bringing structure to the domain. They try to organize sound concepts based on a given criteria, or establish relations among predefined sound categories and sound concepts. For instance, Pierre Schaeffer suggests three viewpoints to describe sound [Schaeffer66]:

- **causal listening**: related to the recognition or identification of physical sound sources
- **semantic listening:** focused on describing the message that is conveyed in the sound, i.e., its meaning
- **reduced listening**: related to the inherent characteristics of a sound regardless its cause and meaning, often using morphological sound criteria.

Several sound taxonomies exist that can be more or less associated to the previous points of view. For example, Peeters suggested a morphological sound description criteria based on the work of Schaeffer [Peeters10], using sound concepts that describe dynamic or melodic profiles, among others. In early 90's, Gaver proposed a taxonomy where sounds are organized considering broad classes of materials and the interactions between them that generate sound. This leads to a limited list of sound concepts related to vibrating solids, aerodynamic sounds and liquid sounds [Gaver93]. Pearce et al. identifies the timbral attributes that end-users commonly search for in online sound effect libraries, covering timbral characteristics of source types and modifiers like hardness, depth, or brightness [Pearce17]. In recent years, other published taxonomies focus more on the *causal listening* perspective, e.g., [Salamon14], where a taxonomy is presented for the specific case of urban sound research, using categories like jackhammer or car horn.

One thing that the aforementioned taxonomies and sound schemes have in common is the specialization in a subset of sound aspects, concepts or categories. Hence, their usefulness from the perspective of reusing Creative Commons content is limited by its application domain, i.e., each of them will apply to a specific subset of sounds.

Last year, Google Machine Perception Team released the AudioSet Ontology[1]: a hierarchical collection of over 600 categories of everyday sounds. Unlike previous taxonomies, AudioSet consists of a

[1] https://research.google.com/audioset/ontology/index.html

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

general-purpose organization of sound. It attempts to cover a very wide range of sound categories and concepts of everyday sounds.

We adopt the AudioSet Ontology for the proposed audio annotator. It is considered a more appropriate resource for the annotation of non-musical content for two main reasons:

- First, it predominantly focuses on the casual listening perspective of sound, i.e., identification of physical sound sources. Given the high diversity of non-musical audio content, it seems reasonable that the emphasis be put on this type of attributes, rather than on other perceptual or morphological aspects that will be applicable only to specific subset of sounds.
- Second, the AudioSet Ontology is, to date, the most comprehensive, i.e., it covers more content types than any other proposed sound organization scheme. This increases the amount of non-musical content that can be described and consequently reused in the context of creative workflows.

1.2 The AudioSet Ontology

The AudioSet Ontology is a hierarchically structured collection of 632 categories of everyday sounds [Gemmeke17]. It is designed to be of general-purpose, covering the acoustic distinctions that can be made by a 'typical' listener, ranging from human and animal sounds to music or sounds made by things. Figure 1 shows the top two levels of the hierarchy.

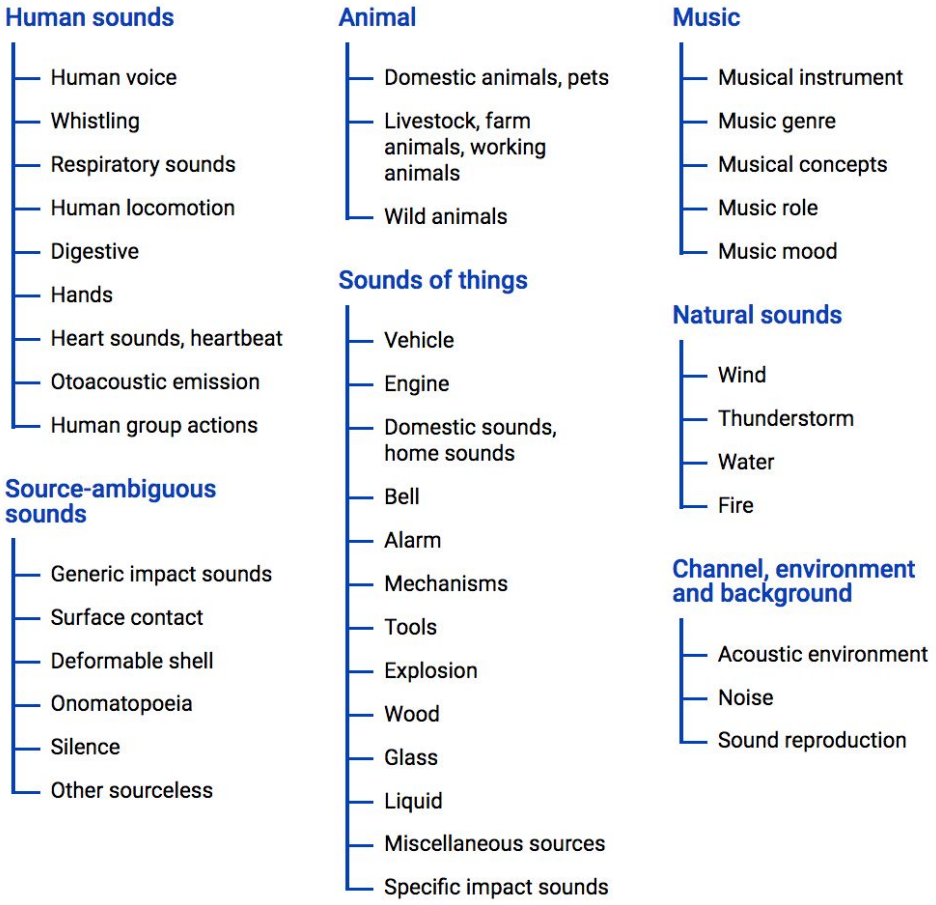

Figure 1. The top two levels of the hierarchy in the AudioSet Ontology.

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

The creation of the ontology started from a large-scale analysis of web text, followed by a series of natural language processing techniques that led to a sorted list of terms. These terms were manually assembled into the hierarchy's first version, which was subsequently refined by comparing it with previous taxonomies and incorporating some of the missing categories. Finally, this category set was used for annotation purposes, and the feedback gathered in this process was utilized to make the modifications that led to the AudioSet Ontology as we know it. The ontology is released as a JSON file featuring the following fields for each category: id, name, description, examples, children and restrictions.

The hierarchical structure has a maximum depth of 6 levels with varying degree of specificity, starting from the general concepts shown in Figure 1 (e.g., Human sounds or Animal) up to other much more specific sound categories that appear at the bottom (e.g., Baby laughter or Bark). The hierarchical structure facilitates the exploration of the sound concepts in a gradual level of complexity, which in turn eases the annotation process.

Another aspect of the hierarchical structure is that it is not a strict hierarchy, that is, there are sound categories that are repeated across several hierarchy paths. For instance, the category Snap occurs in the following two paths:

```
Sounds of things > Wood > Snap
Source-ambiguous sounds > Onomatopoeia > Snap
```

This happens in 33 categories, with four the maximum number of repetitions per category.

One key factor of the AudioSet Ontology is the short descriptions provided per category. Often based on Wikipedia or WordNet (including appropriate citation URIs), they intend to explain the meaning and limit the scope of the category. Although AudioSet Ontology was designed so that the individual categories could be distinguishable by a typical listener, the truth is that not all of them have the same complexity. Moreover, the level of understanding of a given category by a human annotator may easily depend on the cultural and/or academic background as well as the annotator's mother tongue. We think the provided descriptions are key in the context of human labeling in the two main cases that we envision: i) during the process of publishing audio content in the Audio Commons Ecosystem (e.g., when content creators upload content to Freesound), and ii) as a post-annotation step in which users of a platform can collaboratively contribute to the creation of audio datasets (by annotating content for ground truth generation).

In this respect, both hierarchy paths and provided descriptions help to mitigate the potential mismatch between annotators' expertise and hence make the quality of the annotations more uniform. As an example, Table 1 shows two categories that are not particularly simple for a non-native English speaker, and how the addition of paths and descriptions can help the annotator to understand the category.

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

Information type	Category: 'Gurgling'	Category: 'Gargling'
Name	Gurgling	Gargling
Path	`Ontology > Natural sound > Water > Gurgling`	`Ontology > Human sounds > Digestive > Gargling`
Description	The bubbling sound of water flowing through a narrow constriction, such as from a bottle with a narrow neck.	Air from the lungs bubbling through a liquid in the mouth.

Table 1. Two sound categories of the AudioSet Ontology that may not be simple to understand for non-native English speakers. The addition of paths and descriptions clarifies their meaning and scope.

As mentioned, the AudioSet Ontology is the most comprehensive set of sound-related categories to date, including 632 categories. Because of this, the nature of the categories included varies to a high extent. Table 2 lists our view about the different types of categories, along with an explanation and a few examples[2].

Category Type	Explanation	Examples
Sound events	Sound produced by a physical sound source, with a defined duration in time [Virtanen18]	'Dog bark', 'Jackhammer'
Acoustic scenes or contexts	Entirety of sound formed by the mixture of various typical sources from a real scenario [Virtanen18]	'Inside, small room' or 'Outside, rural or natural''
Production mechanisms	E.g, deformation or impact of materials that otherwise would radiate no sound	'Tearing', 'Splinter', 'Fill (with liquid)'
Properties of sound	Categories that describe aspects of sound	'Reverberation', 'Echo', 'Cacophony', 'Whir', 'Boing'
Arrangements of other sound concepts	Consisting of specific arrangements of other existing categories	'Clatter', 'Applause', 'Narration, Monologue'
Mathematical signals	Signals that are typically defined mathematically	'Sine wave', 'Chirp tone', 'Pink noise'
Abstract categories	*Portmanteau* categories that exist to give structure to the ontology (not to be used as labels)	'Onomatopoeia', 'Human locomotion', 'Digestive'

Table 2. List of category types included in the AudioSet Ontology.

[2] Note that the listed category types are not mutually exclusive as their separation is highly subjective.

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

In the AudioSet Ontology, sound events and production mechanisms are the most predominant category types. On the contrary, categories describing aspects of sounds such as the timbral attributes defined in [Pearce17] are less abundant. In this respect, the ontology could be further expanded in future iterations. The ontology also covers a number of music-related categories, as can be seen in the top two levels of the hierarchy in Figure 1, although that is not the focus for the sake of the proposed annotator.

1.3 Challenges of large vocabulary audio annotation

Adopting the AudioSet Ontology for annotation of non-musical sounds yields a number of challenges. Next, we summarize the main challenges based on the discussion elaborated in previous sections.

- **Large vocabulary.** Users have to select among 632 sound categories (roughly 450 belong to non-musical content). Efficient exploration of them is essential.
- **Multiple aspects to annotate**. Categories under consideration are of very different nature (see Table 2).
- **Vocabulary not necessarily known by all annotators**. The understanding of every category will depend on cultural and/or academic factors as well as annotator's mother tongue.
- **Semantic overlap**. For a determined acoustic content, several categories may apply. For instance, the sound of an electric toothbrush should be labeled as 'Electric toothbrush', 'Engine' and 'Whir'. Having an exhaustive annotation of the content will ease its search and reuse.
- **Varying level of specificity**. While some categories are rather generic (e.g., 'Human sounds') others are very specific (e.g., 'Baby laughter').
- **Multiple parents.** The ontology is not a strict hierarchy, as some categories occur more than once in different places (i.e., contexts) of the ontology. For instance, 'Hiss' is a son of four categories: 'Steam', 'Cat, 'Snake' and 'Onomatopoeia'.

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

2 The Audio Commons Annotator

The Audio Commons Annotator is a web based tool for the manual annotation of non-musical content. It provides an interface which intelligently guides the user in the process of annotating audio samples with a wide range of categories. The interface is mainly developed for being used during the process of publishing audio content in the Audio Commons Ecosystem (e.g., when content creators upload content to Freesound). Nevertheless, it can also be used as a post-annotation step in which users of a platform can collaboratively contribute to the annotation of content and ground truth generation.

The prototype developed is not integrated in any content provider platform yet, but it is implemented in a way that it can be easily be done. Moreover, we implemented it in the context of Freesound Datasets,[3] a platform for the collaborative creation of open audio collections labeled by humans and based on Freesound content. In this platform, we took advantage of some available tools and audio content to propose features such as additional pop-up informations as well as exploratory tables for guiding users in the annotation process. The tool instance implemented in Freesound Datasets is the one that we present in detail in this deliverable. Some of its functionalities are tied to the specific use case of the AudioSet Ontology.

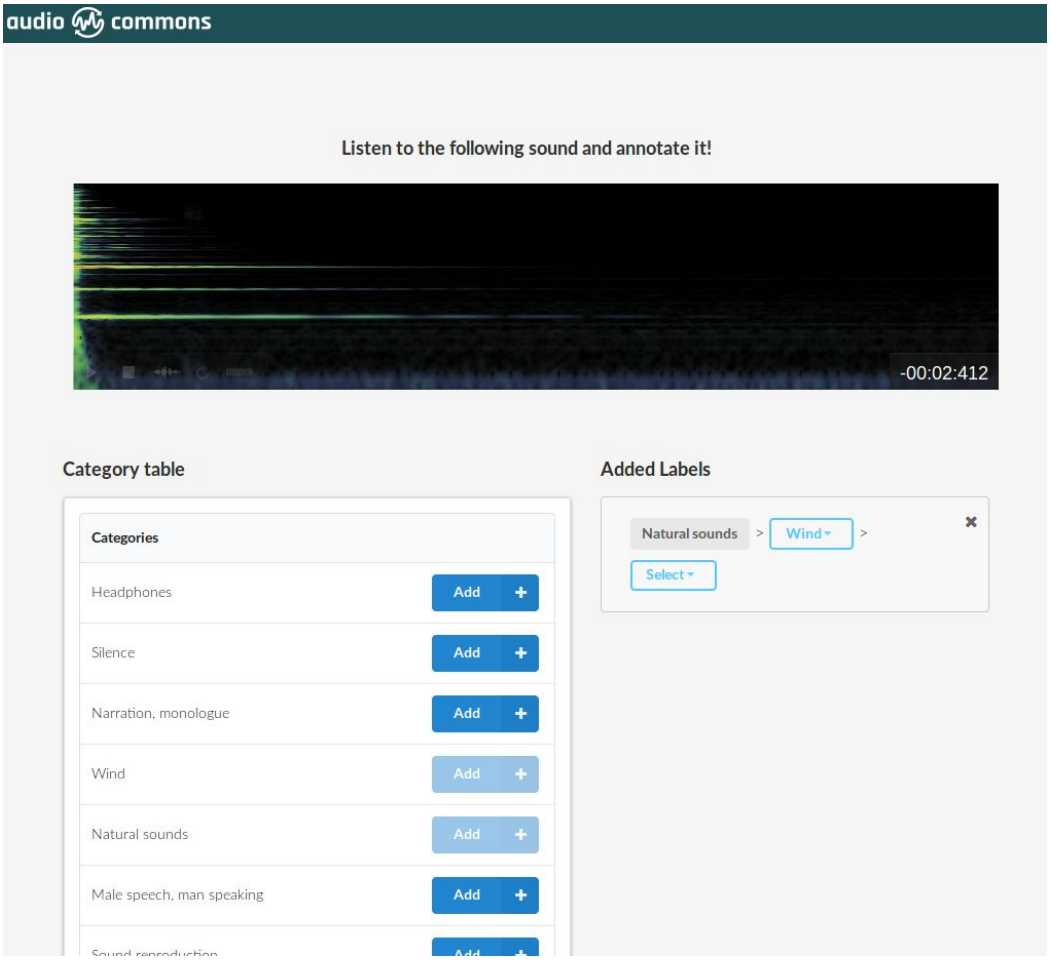

Figure 2. Screenshot of the annotator interface

[3] 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

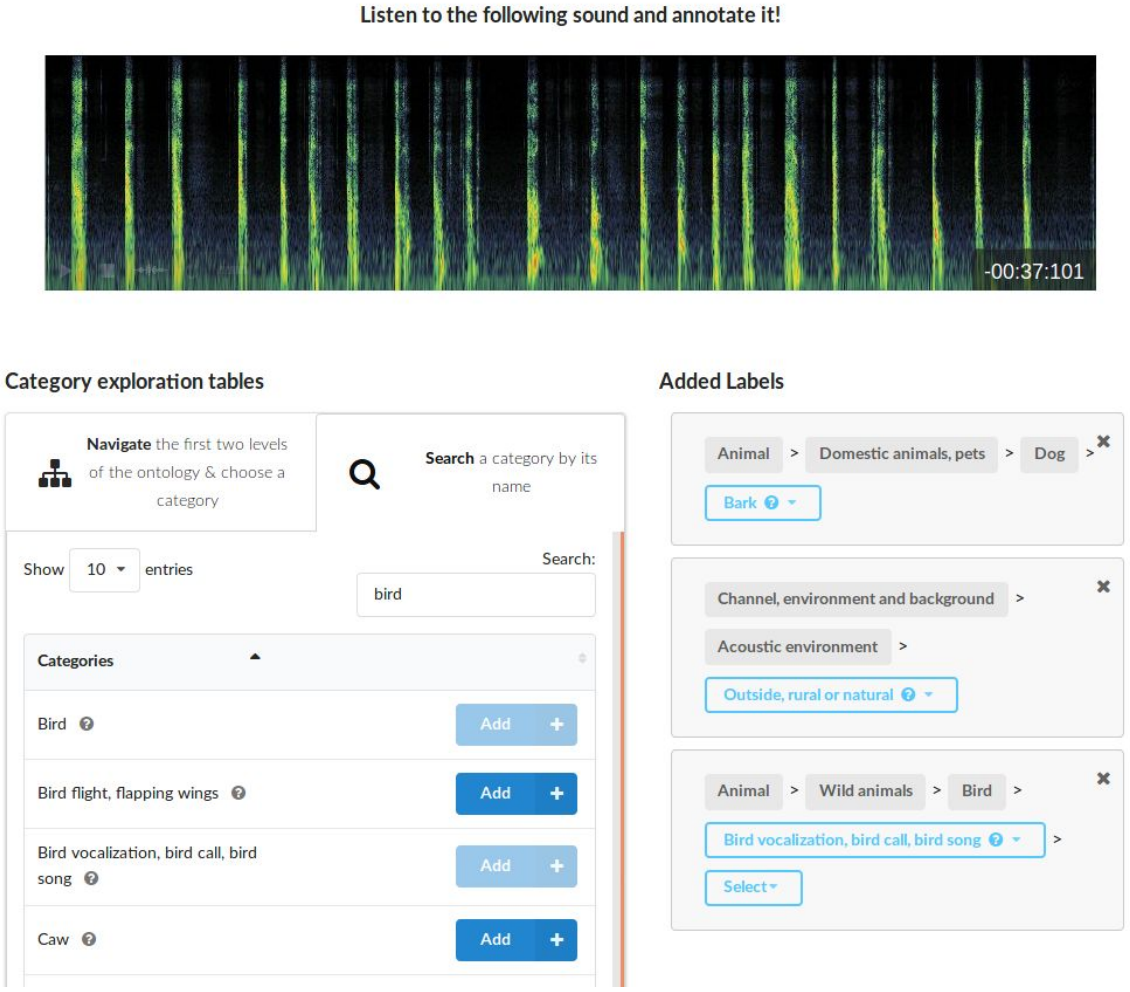

Figure 3. Screenshot of the annotator interface implemented in the Freesound Datasets platform.

2.1 Concept and main functionalities

The main idea of the interface is to intelligently guide the user through the process of annotating audio content. As shown in the previous section, assigning labels from a large vocabulary to audio resources is a difficult task for non-experts and creates a number of challenges that we try to mitigate. To ensure high quality annotations, we propose an interface that combines two different approaches: (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.

2.1.1 Exploration stage

In the Freesound Datasets platform implementation, we propose two tables for enabling the exploration of the taxonomy. Both tables include pop-up showing descriptions and audio examples of the sound categories.

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

Navigation taxonomy table

In the first table, it is possible to visualise and navigate through the first two levels of the taxonomy, and finally select some labels. This aspect of the tool intends to provide efficient exploration, as a user can rapidly scan the different labels available in the taxonomy. Moreover, exploring the taxonomy by topic may help the user discover terms that she could otherwise have missed. Figure 4 shows the first level of the AudioSet Ontology, along with the description provided for one category. Figure 5 shows an end level as a result of selecting 'Animal' and 'Livestock, farm animals, working animals' categories.

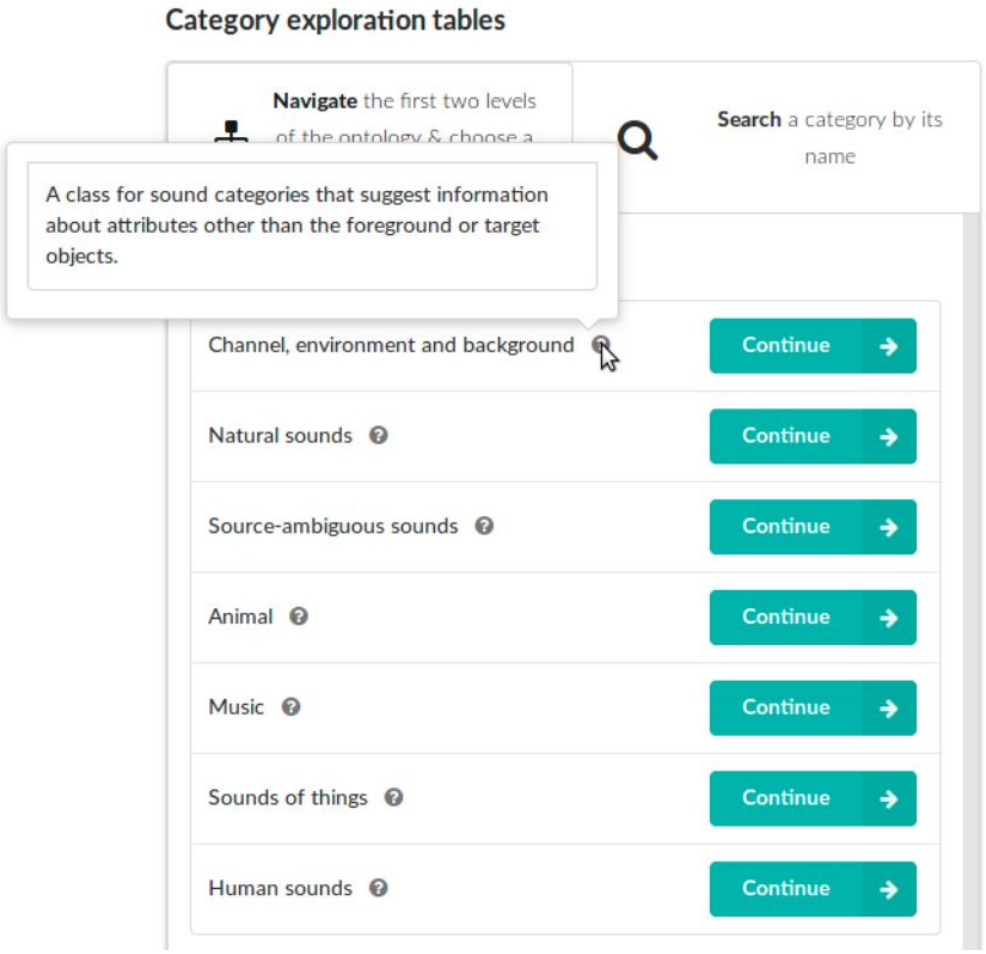

Figure 4. Screenshot of the first level of the navigation taxonomy table showing a pop-up description.

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

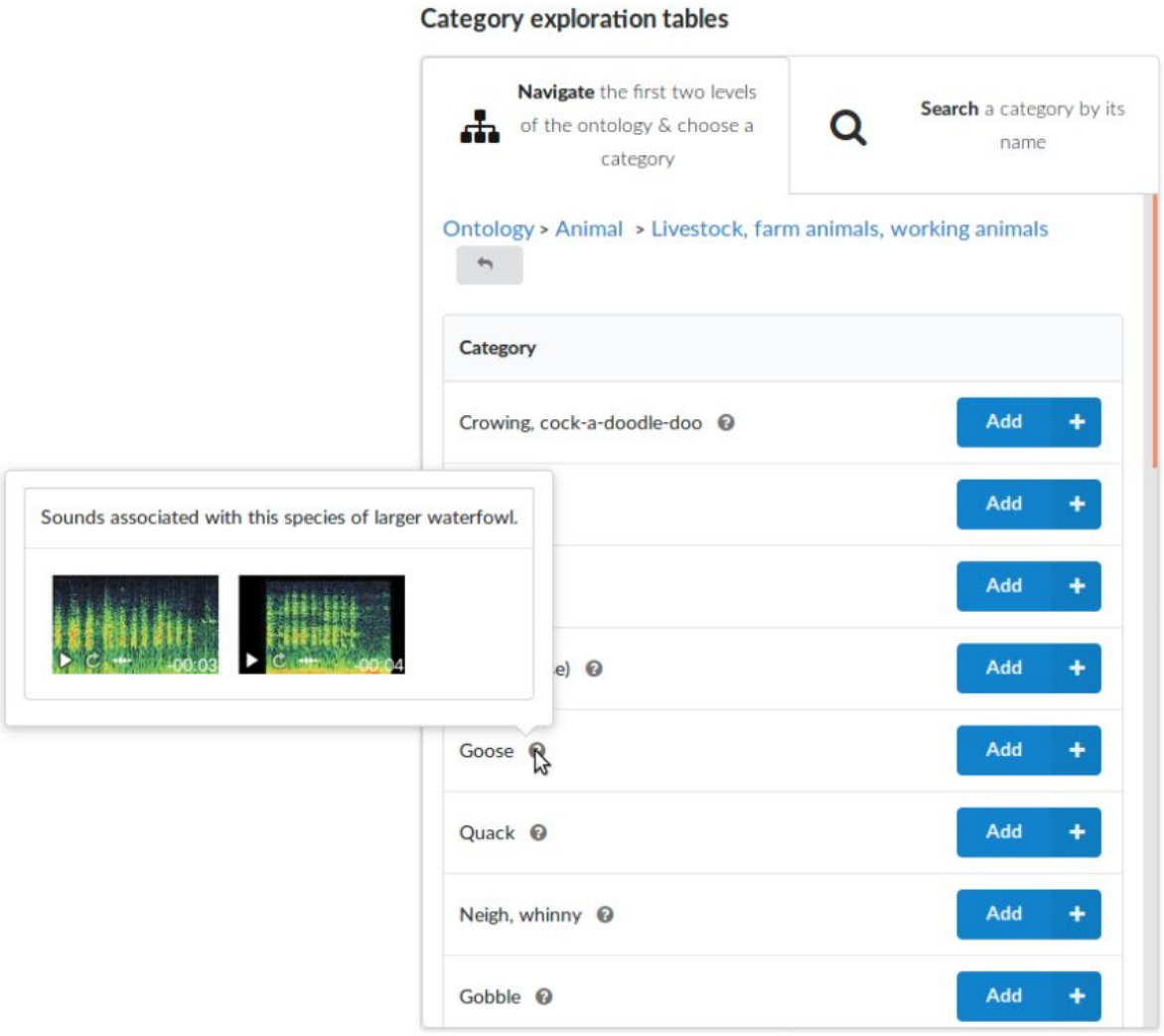

Figure 5. Screenshot of the lowest level of the navigation taxonomy table in the which a user can add some labels.

Search taxonomy table

The second table allows (i) to explore all the labels in a flat representation, and (ii) to perform text queries which will retrieve categories that have the query terms in their name, description or hierarchy path(s). This can allow users lacking some vocabulary to find a label using their own words if included in its description. Figure 6 shows the result for the query “animal”.

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

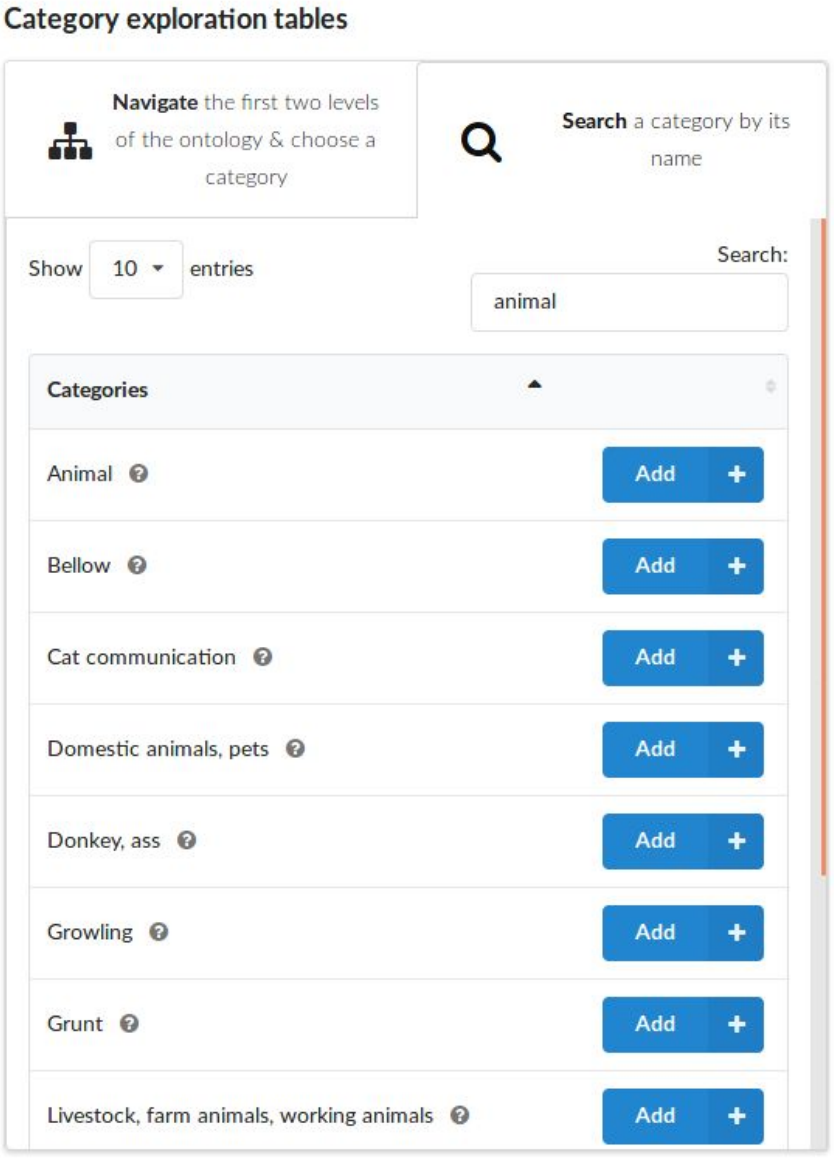

Figure 6. Screenshot of the search taxonomy table with the query “animal” where users can directly add labels.

2.1.2 Refinement stage

In a second stage, after having added one or several labels, the user is presented the added labels with their hierarchy paths in the ontology. The motivation for introducing this refinement stage is to ensure better quality annotations by allowing users to visualize the whole path of the added label, and also providing information about related categories that might be more suitable for describing the content being annotated.

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

Sibling categories

As shown in Figure 7, users can modify the added label(s) by selecting a sibling category from a dropdown input. This intends to provide a complementary tool for exploring categories from the taxonomy.

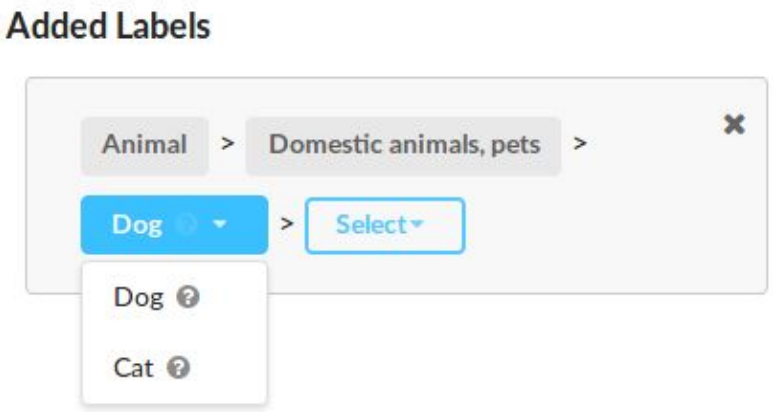

Figure 7. Screenshot of the 'Dog' label at the refinement stage. The dropdown shows the sibling categories.

Child categories

Another challenge we need to adress is the fact that the content needs to be annotated as precisely as possible, using very specific categories. The hierarchy of the AudioSet Ontology implies that those at the bottom provide a more complete and precise description of the content. For instance, in Figure 8, the interface highlights the fact that there are more specific categories than Dog. The user can explore and specify a more precise label from a dropdown listing the children labels.

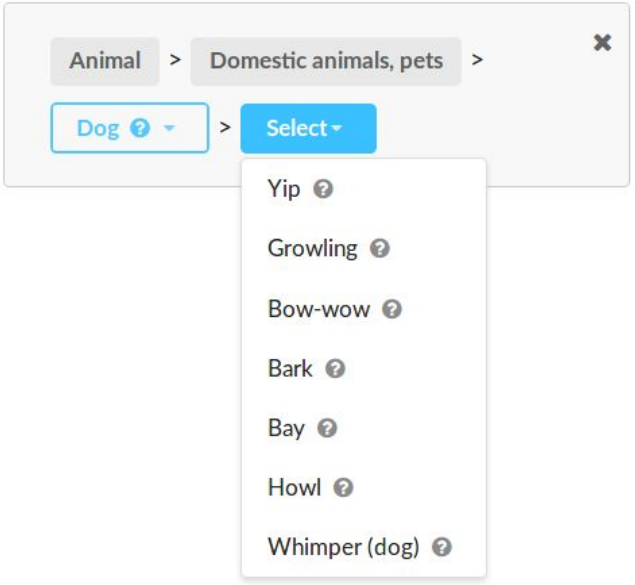

Figure 8. Screenshot of the 'Dog' label at the refinement stage. The interface shows existing child categories which are more specific than the one previously added.

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

Multiple parents

Another characteristic of the AudioSet Ontology is that it is not a strict hierarchy as explained in Section 1.2. A category can have multiple parents and thus have different meanings. To address this ambiguity problem, we add all the corresponding paths as different labels when a user selects a label during the exploration stage. Figure 9 shows the added labels when selecting 'Hiss'. The user should delete the paths that do not correspond to the content being annotated.

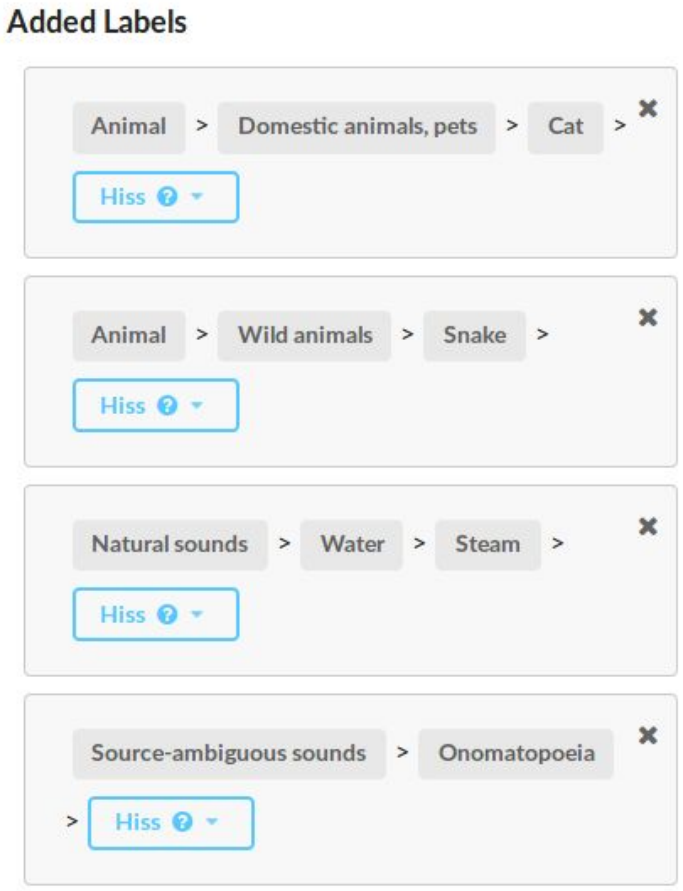

Figure 9. Screenshot of the added labels when selecting Hiss from a taxonomy table.

2.2 Implementation

The Audio Commons Annotator prototype demonstrator is implemented using the Python programming language and the popular Django[4] web framework. The source code is open source, released under Apache 2 license and available at https://github.com/AudioCommons/ac-annotator in the *annotator* project application. Except for the exploration tables and the pop-ups, the interface is fully implemented in JavaScript, which makes it suited for its integration in web applications.

[4] https://www.djangoproject.com/

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

The easiest way to set up a local development version is to use Docker[5] container definitions provided in this repository. The instructions for setting up the development environment using Docker are the following:

- Before starting, make sure you have Docker (with `docker-compose`) installed.
- Clone repository and cd into it:
  >>> git clone git@github.com:AudioCommons/ac-annotator.git
  >>> cd ac_annotator
- Build and run Docker containers for required services:
  >>> docker-compose up
- Now you should be able to access your server at `http://localhost:8000`

The current prototype is accessible at http://localhost:8000/generate_annotations/

2.2.1 Data structure and label group representation

The hierarchical taxonomy is sent to clients from an asynchronous AJAX request. To avoid complicated tree structure manipulation on the client side, the hierarchical structure is transformed into a list of category sequences (all the possible paths from a top category to a bottom category). For the AudioSet Ontology, it corresponds to 526 different arrays (e.g. ["/t/dd00098", "/m/05n1m", "/m/07rjwbb"], which correspond to ["Source-ambiguous sounds", "Onomatopoeia", "Hiss"][6]). Note that the AudioSet hierarchical taxonomy is not a mathematical tree structure but rather a directed acyclic graph, as some nodes have multiple parents. This is also one of the reasons why we avoided to send the structure as it is and prefered to send all the possible paths in the hierarchy. Each path is a different concept whereas a label in itself can be associated to multiple concepts (e.g., 'Hiss' accepts 'Cat', 'Snake', 'Steam' and 'Onomatopoeia' as parents). In the Audio Commons annotator repository, we added as an example:

- A hierarchy paths list taken from the AudioSet Ontology
- The corresponding category names

When adding a label from a taxonomy table, one or several group(s) of labels, corresponding to the selected label hierarchy path(s) is/are added. In a label group, the selected label is included in a dropdown input which contains also its siblings in the taxonomy. If more specific child labels exist, another dropdown is appended containing the corresponding children.

2.2.2 Integration into web applications

In the Audio Commons prototype, a Django template is used to create the HTML page that contains the main elements for allowing the Javascript to dynamically add the label groups.

The `generate_annotations` view function from `ac-annotator/annotator/views.py` renders the `ac-annotator/annotator/generate_annotations.html` template and sends a Freesound sound ID as an example. The HTML template also contains the JavaScript code. The other two functions in the `views.py` file (`get_hierachy_paths()` and `taxonomy_table()`) are serving the AJAX requests (one for sending the hierarchy paths and category names, and the other one for sending the taxonomy table content). Note that the structure is duplicated there, as these

[5] https://www.docker.com/products/overview
[6] The category identifier strings are the one initially used in the AudioSet Ontology.

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

informations might come from processing of data coming from one source file or from a database (like in Freesound Datasets).

In the current prototype, there is no possibility to submit the generated labels. However, the JavaScript `getAddedLabels()` function collects all the added labels in the page. For instance, an AJAX POST request could be used to send the results to the server.

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

3 Conclusion

In this deliverable, after having identified the challenges when annotating non-musical content with large vocabularies of audio related concepts, we described an initial version of a prototype tool for the manual annotation of non-musical content. At the moment, the prototype consists of a web-based interface that intelligently guides users in the annotation process. First, the user explores and selects labels, and afterwards she visualizes and refines them.

The prototype developed here will continue evolving during the next phases of the project. Our plan is to progressively incorporate new features such as the inclusion of recommended labels. For instance, after having added one label, the system could recommend semantically overlapping labels. This would help users to fully annotate their content by adding all the categories that apply.

Another aspect that we overlooked and deserves attention, is providing clear and complete guidelines during the annotation process. Specific questions should be asked to the user, and ideally they should guide him to focus on the multiple aspects that need to be addressed. The category types proposed in Table 2 suggest that users could be asked about source identity information or production mechanisms, acoustic contexts and properties of the sounds.

Finally, we may consider allowing users to leverage their own words for specifying some of the audio related concepts that might be missing in the adopted taxonomy. This would be useful for instance to extend the Audio Commons Ontology with labels that were not initially included. Also, it could provide hints to what categories are not well understood or found by users during the *exploration stage*.

As part of this task, an evaluation of the interface will be carried out in terms of its usability and its expressive power for annotating non-musical content. This evaluation will be carried out with real users in the context of Freesound Datasets in Task 5.5.

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

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.

[Gaver93] Gaver, William W (1993). “What in the world do we hear?: An ecological approach to auditory event perception”. In: Ecological psychology 5.1, pp. 1–29.

[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.

[Pearce17] Pearce, Andy; Brookes, Tim; Mason, Russell (2017) “Timbral Attributes for Sound Effect Library Searching”. In: AES International Conference on Semantic Audio

[Peeters10] Peeters, Geoffroy and Emmanuel Deruty (2010). “Sound indexing using morphological description”. In: IEEE Transactions on audio, speech, and language processing 18.3, pp. 675–687.

[Salamon14] Salamon, Justin, Christopher Jacoby, and Juan Pablo Bello (2014). “A dataset and taxonomy for urban sound research”. In: Proceedings of the 22nd ACM international conference on Multimedia. ACM, pp. 1041–1044.

[Schaeffer66] Schaeffer, Pierre (1966). “Traite des objets musicaux”. In: Paris: Seuil.

[Virtanen18] Virtanen, T., Plumbley, M.D. and Ellis, D. eds., (2018). “Computational Analysis of Sound Scenes and Events”. Springer.

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