(GIS) OSmOSE

Open Science meets
Ocean Sound Explorers

APLOSE: a scalable web-based annotation tool for marine bioacoustics

OSmOSE Product Presentation

Authorship This document was drafted by

- Paul Nguyen Hong Duc[1)]
- Maëlle Torterotot [2)]
- Romain Vovard[3)]
- Erwan Keribin[4)]
- Dorian Cazau[5)]

belonging to the following institutes (at the time of their contribution): 1) Sorbonne Universités, 2) IUEM, Université de Brest, 3) Freelance developer at Élan Créateur, 4) Freelance developer, 5) Lab-STICC, ENSTA Bretagne.

Document Review Though the views in this document are those of the authors, it was reviewed by a panel of acousticians before publication. This enabled a degree of consensus to be developed with regard to the contents, although complete unanimity of opinion is inevitably difficult to achieve. Note that the members of the review panel and their employing organisations have no liability for the contents of this document.

The Review Panel consisted of the following experts (listed in alphabetical order):

- Ronan Fablet[1)]

belonging to the following organisms / research institutes (at the time of their contribution): 1) Lab-STICC, IMT Atlantique.

Last date of modifications June 9, 2020

Recommended citation Nguyen, P. et al. "APLOSE: a scalable web-based annotation tool for marine bioacoustics", OSmOSE Product Presentation (version dating from June 9, 2020, distributed openly on `https://osmose.xyz/`)

Future revisions Revisions to this document will be considered at any time, as well as suggestions for additional material or modifications to existing material, and should be communicated to Dorian Cazau (dorian.cazau@ensta-bretagne.fr).

Acknowledgements We thank the Pôle de Calcul et de Données pour la Mer[1] from IFREMER for the provision of their infrastructure DATARMOR and associated services. We also would like to thank our main sponsors in this work: CominLabs[2] through the innovation action Tech4Whales, DREC Agence Française de la Biodiversité[3] and ISblue[4]. The authors also would like to acknowledge the assistance of the review panel, and the many people who volunteered valuable comments on the draft at the consultation phase.

[1] `https://wwz.ifremer.fr/Recherche/Infrastructures-de-recherche/Infrastructures-numeriques/Pole-de-Calcul-et-de-Donnees-pour-la-Mer`

[2] `https://www.cominlabs.u-bretagneloire.fr/`

[3] `https://www.afbiodiversite.fr/`

[4] `https://www.isblue.fr/about-us/`

Contents

Abstract

Lately, underwater passive acoustics has been used extensively for a variety of purposes, from monitoring of marine mammals populations to surveillance of human activities. Recent technological improvements have made it possible to increase the frequency range and autonomy of acoustic recorders, resulting in an enormous amount of data. Manual inspection of these huge datasets is unmanageable and a need for automated recognition, detection and classification methods has emerged. However, in order to train and test the performance of these algorithms, data subsets must still be annotated by human operators. To date, the difficulty of creating large amounts of accurately annotated data has been a major obstacle to build accurate underwater sound recognition algorithms.

In this paper, we present APLOSE, an open-source, web-based yet scalable tool which should highly facilitate collaborative annotation campaigns in marine bioacoustics. This would eventually lead to the creation of reference datasets that could be used to build robust, low-bias detection and classification algorithms.

Chapter 1

Introduction

1.1 Context

A variety of animals produce species-specific acoustic signals, including marine mammals (Richardson et al., 1995), fish (Anorim, 2006), crustaceans (Versluis et al., 2006)... Acoustic analysis has become a standard method in studies of animal vocal communication, and manual detection of acoustic cues was initially the common practice. However, advances in recording hardware speeds, battery life and data storage capacity have increased the rate of acoustic data accumulation to a point where reliance on manual analysis has become unmanageable.

Automated detection and classification algorithms have become necessary for the analysis process. These algorithms provide more consistent and comparable estimates throughout a study period and across studies when processing long-term time series. They are less prone to bias than human analysts, and can be quantified more objectively. However, they cannot be used without supervision, and typically require performance evaluation or correction at some point in the processing pipeline. For instance, labelled datasets are used for training and evaluation of machine learning models, and misclassifications may need to be quantified or corrected (Marques, 2013). Manual review remains also an important part of the process for additional scientific insights, since analysts are best able to judge the context-dependent nature of biological data. Such a supervision often goes through a manual annotation process by one or several analysts. Obtaining such expert annotations is resource intensive and laborious, especially for long recording campaigns as a reliable annotation often needs a careful listening of sounds. Overall, either the lack or quality level of annotated datasets are now frequently criticized (e.g. Leroy (2018), see last slide from `http://cetus.ucsd.edu/dclde/docs/pdfs/Wednesday/14-Gillespie.pdf`), preventing our community to comply with the best practices in machine learning development, with for example the construction of sustainable reference benchmarking datasets.

1.2 Motivations and objectives

In its essence, our project fulfills the general objective of developing an open source and collaborative tool for annotating long-term passive acoustic data, underpinned by the motivation of increasing annotation result and method sharing and visibility. By encouraging a data sharing culture amongst the passive acoustics community, as well as facilitating data discussion and multi-user support, such collaborative environment like APLOSE should help performing better research in less time (Lowndes et al., 2017).

This combination of free and open tools, infrastructure and a collaborative environment has been highlighted as key requirements to lead to more informed conclusions and management decisions, as already observed in similar projects (e.g., Whalenet[1], the Narragansett Bay Coyote StudyWhalenet[2], the Information System for the Analysis and Management of Ungulate Data (Cagnacci and Urbano, 2008)). This

[1] `http://whale.wheelock.edu/Welcome.html`
[2] `http://www.theconservationagency.org/coyote.html`

increased synergy may also reduce the need to collect further or new data and can lead to academic as well as financial gains (Huettmann, 2005; Boulton et al., 2012).

More specifically to our community, at least long-term three research avenues motivated our work. The first general one is to speed up the process of large marine bioacoustics dataset annotation while improving annotation quality, by providing a facilitator tool for collaborative annotation. Indeed, such a collaborative approach can potentially reduce time of a given annotation task by distributing it to multiple annotators, while some overlapping parts of their annotation tasks can be used to cross-validate them. More labels with higher-quality will allow for sure the development of more reliable machine learning algorithms, especially through the creation of larger benchmarking datasets (Piczak, 2015). A second motivation was to be able to generate analytical supports to help understanding annotators' behaviour by capturing the ambiguity that its content might produce in relation to the defined ontology, quantifying individual biases and better characterizing what makes the quality of audio annotations (as already done in other communities like e.g. in urban soundscapes (Cartwright, 2019)). It is known that low-quality human annotations indirectly contribute to the creation of inaccurate or biased learning systems. A third motivation would be to be able to handle large-scale crowd-sourcing campaigns. Crowdsourcing has been shown to be a viable alternative to conventional labelling paradigms to rapidly collect the mass of annotations needed to leverage new data sources. Crowd-sourcing programs increasingly thrive in the marine realm (e.g. Southern Ocean Research Partnership[3], OBSenMER[4], HappyWhale[5]), although it is much less favourable than the terrestrial one where year-round participative surveys can be easily organized, e.g. for birds[6]. Whilst crowdsourcing has many positive aspects including efficiency and cost reduction, the online recruitment of anonymous annotators brings new and different issues especially in relation to the annotation quality.

In line up with these motivations, our development objective was to offer an easy way to annotate complex bioacoustics sound events within dynamic multisource soundscapes, with the requirements of being collaborative, user-friendly, scalable (adapted to large scale datasets), flexible (creation of fitting ontologies for the multiple tasks it can be applied to, easy integration of new apps), as well as open source.

1.3 Related works and contributions

1.3.1 From neighboured communities

To date, a variety of tools have been created for the annotation of audio events, e.g. the famous Praat[7] and HAT[8] for human speech, MUCOSA (P. Herrera and Fabra., 2005) and Sonic Visualizer (C. Cannam and Bello., 2006) for music or more general-purpose tools like ASAnnoatation[9] (Bogaards, 2008) which also provides low-level feature information (e.g. pitch content). These tools allow for a wide variety of analysis, like annotation and generation of music metadata at different abstraction levels including a collaborative annotation subsystem, as well as analysis and feature extraction applications, for MUCOSA. Also, tools to deal with multimedia data have also been developed, such as ELAN (P. Wittenburg and Sloetjes., 2006), a full-featured and complex tool that allows the annotation of both audio and video.

All these tools share the property of being only locally executable. It is only more recently that annotation tools started to get online, particularly impelled by the open-source release of key software components. Among them WaveSurfer (Sjolander and Beskow, 2000)[10], a tool initially designed for speech annotation but deliberately made flexible and extensible to different tasks. Based on it, CrowdCurio[11] is a JavaScript web interface for the annotation of audio events that uses and extends the Regions plugin of wavesurfer.js, including useful features such as labeled regions or the possibility to switch the sound visualization between its waveform and its spectrogram. CrowdCurio has already been adopted in the ISMIR community with Melendez-Catalan et al. (2017)'s BAT interface, allowing the user to label different audio sub-regions with

[3] `http://www.marinemammals.gov.au/sorp/sightings`
[4] `https://www.obsenmer.org/pages/presentation`
[5] `https://happywhale.com/home`
[6] `https://www.citizenscience.gov/ebird-bird-data/#`
[7] http://www.fon.hum.uva.nl/praat/
[8] http://www.speech.kth.se/hat/
[9] `http://recherche.ircam.fr/anasyn/ASAnnotation/`
[10] http://www.speech.kth.se/wavesurfer/
[11] https://github.com/CrowdCurio/audio-annotator

overlapping events by giving a salience rate of the different recognized sounds in that segment. Ontologies and cross-annotation are also featured with this tool.

More advanced machine learning oriented development have also been proposed. I-SED (Kim and Pardo, 2017) is a web-based tool for speech recognition involving a machine learning method to help the labeler spotting audio segments similar to the target sound to find. Then, the annotator decides whether or not accepting them. It helps reducing the annotation time. SoundScape (Krijnders and Andringa, 2009) also uses machine learning, and it allows the annotation of specific time-frequency regions of the spectrogram.

1.3.2 Within the bioacoustics community

In the bioacoustics community, similar development trends can be observed. Highly specialized tools have been developed such as the famous Raven Pro software[12]. It is a prevalent software program for the acquisition, visualization, measurement, and analysis of sounds, used by many bioacousticians to annotate their datasets (e.g. Leroy (2018)). The Pro version (Raven Pro), costs from $50 up to $800 dollars depending on the license type and terms. It allows the user to visualise the sound as a waveform and/or a spectrogram. Multiple parameters may be modified by the user, such as the spectrogram windows' sizes, the contrast, the colorbar. Presets can be created, saved and downloaded with custom parameters. The user can also zoom in and out and play the sound at different rates. The annotations are made by drawing a time/frequency box (or time only if the annotations are done on the waveform) around the acoustic event of interest. Annotation are then stored in an 'Annotation Table' with the start and end time and upper and higher frequency values of each "box". Multiple annotation tables can be filled at the same time when annotation is performed on numerous call types simultaneously. The Lite version (Raven Lite) is free and provides the basic functions of Raven Pro, but don't allow for advanced and customizable control of spectrogram parameters and advanced sound measurements and annotation. A more generic open source software is Audacity[13], offering also many features for sound annotation. The user needs the dataset in local to be able to perform audio transformation and annotation.

Like for the other communities, these softwares are not distributed as online interfaces, impeding noticeably collaborative works on a same dataset. Only recently first web-based appear, such as Koe[14], a web-based application for classifying and analysing animal vocalizations (Fukuzawa, 2019). Koe offers bulk-labelling of units via interactive ordination plots and unit tables, as well as visualization and playback, segmentation, measurement, data filtering/exporting and new tools for analysing repertoire and sequence structure - in an integrated environment.

1.3.3 Contributions

Following these efforts, our work aims to provide a new web-based annotation interface dedicated to marine bioacoustics datasets, able to answer some limitations of existing tools in our community. In table 1.1, we list the main non-functional requirements of APLOSE, and below we further discuss how they are addressed by these tools. Note that we left aside more functional features that are generally shared in the system design of these tools, such as exploring the data interactively using smooth zoom and scrolling, providing time/spectral annotation box, spectrogram contrast, audio player with speed variation, user-defined labels... Many of these features were not present on the original CrowdCurio version on which APLOSE is based, although they are crucial to allow for a more flexible annotation practice. Those will be further described and illustrated for APLOSE in Sec. 2.1.

Accessibility

In this first and perhaps most important non-functional requirement, we qualify the capacity of an annotation system to "easily integrate any potential campaign participant". Any web-based technology like APLOSE or Koe will perfectly answer this requirement as it simply comes down to the need of having a web browser on its computer. This should be highly beneficial as sharing of data or granting access to a web application is

[12] `http://ravensoundsoftware.com/software/raven-pro/`

[13] `https://www.audacityteam.org/`, see manual at `https://manual.audacityteam.org/man/spectrogram_view.html`

[14] `https://koe.io.ac.nz/#!`

much easier than sharing data directly, and then avoids the situation where every annotator has to manipulate/install data and tools locally, which may raise complex logistic problems, even more with high-volume datasets. More anecdotally maybe, this requirement will facilitate annotation to anyone working across multiple computers and devices, or not having local administrative access on their computer, which may also become tricky when dealing with proprietary and non-inter-operable softwares.

Scalability

Here, we qualify the capacity of an annotation system to "easily integrate a high number of campaign participants, even if that number shall increase during the campaign". In other words, scalability is the property of a system to handle a growing amount of work by adding resources to the system (Bondi, 2000), and globally to facilitate performance requirement regardless the work load. One dimension (sometimes referred to as **administrative scalability**) on which measuring it is the user ability to access the system for an increasing number of organizations or users. Then, an effective scalability allows to enlarge the previous requirement of accessibility to a very large number of annotators even during a given annotation campaign. This second requirement now allows for the creation of real-time collaborative environment where multiple users can access the same project and can interact with each other.

One of the biggest most valuable and predominant features of modern cloud-based services, as used by APLOSE and Koe, is simplified scalability, being often the primary requirement of IT environments. Especially, horizontal scaling, delivering both performance along with storage capacity, allows a total workload volume to be aggregated over the total number of nodes and latency is effectively reduced.

Performance

Through the performance requirement, we qualify the capacity of an annotation system to "offer a good user experience, even if that experience shall involve high-volume datasets". To have a good user experience the response time of the user interface has to be less then a second, over which users tend to consider them as waiting time[15]. As we shall see in further details in Sec. 3.1, performance becomes critical for certain annotation use cases that need a higher volume of annotation data per displayed window, like for long-term scene classification or context-based event annotation, as waiting times tend to drastically increase with the size of the displayed duration, up to getting too high for a satisfying user experience. For example, a sound file as distributed in the last DCLDE2015HF workshop will take on average more than 20 seconds to be opened on audacity using a standard professional laptop. Here performance can be linked to the notion of **computational scalability** which is the ability of processing more data in a reasonable amount of time. Put in other terms, the processing for visualization purposes should be able to handle large amount of acoustic data with the same performance as for smaller dataset.

APLOSE was developed to be able to handle larger volumes of annotation data than currently done without scarifying performance, thanks to two modern web-based technologies, namely cloud-based services and tile-based rasterization, similarly to the Google Pattern Radio system[16]. First, as already mentioned, cloud-based offers a simplified horizontal scaling, which is ideal for workloads that require reduced latency and optimized throughput as for annotation interface. Koe shares with APLOSE this capacity, while other tested softwares are not cloud-based. Second, as we describe in more details in Sec. 2.2.3, the tiling technology used in APLOSE provides an efficient solution for annotation practices requiring higher volumes of data than currently done, as all spectrograms are pre-computed before the annotation campaign and serve to the client as compressed images of smaller duration tiles (i.e. segments). On the contrary, Koe performs a client-side spectrogram computation which highly limits the processing and rendering of large annotation datasets. But note that this way, Koe achieves full interactivity with browsed data using vector data directly in the browser.

Extensibility

Through the extensibility requirement, we qualify the capacity of an annotation system to "be easily extended to meet new needs while inducing a minimal level of efforts to implement the extension". There are many different dimensions behind this notion, such as cost-effectiveness in terms of both software licensing and

[15] See for example `https://www.nngroup.com/articles/response-times-3-important-limits/` for more details.
[16] `https://patternradio.withgoogle.com/`

hardware. Ideally, a system should be freely distributed with copyleft licenses which guarantee that future versions of the software will remain free and publicly available. This is an important factor for smaller research bodies, consultancies and conservation organisations who have limited resources to dedicate towards software purchasing (Tufto and Cavallini, 2005). Concerning hardware, extending an annotation system so it can meet increasing demands on the number of annotators or volume of data to be processed can induce severe extra-cost if the system architecture has not integrated such extension from scratch.

Once again, one major benefits of cloud computing, as used by APLOSE and Koe, is precisely cost-effectiveness, as it allows a processing activity to grow without making any expensive changes in the current setup thanks to scalability in the cloud, reducing significantly the cost and effort implications of storage and processing resource growth, compared with hosting the system locally. In practical terms, for APLOSE, we just have to commission additional virtual machines to scale out to a larger amount of annotators or data. For what concerns free and open source distribution, as for BAT (Melendez-Catalan et al., 2017), our software development is based on open source software components like CrowdCurio. This is shared by most annotation tools distributed (at least) in the research community, excepting Raven Pro X (although a free version exists).

With a more programming language point of view, extensions can pass through the addition of new functionalities or through the modification of existing ones without impairing existing system functions[17]. Such extensions are then highly facilitated by the use of high-level programming languages like Python and Javascript instead of low-level ones like C++, making them more user-friendly for less technically skilled users. To maximise software sustainability and extensibility, in APLOSE we exclusively used such high-level programming frameworks, and no additional software, add-on packages or plugins are required to run, visualise or export the raw, filtered or analysed data other than the Web browser (e.g. Edge, Firefox and Chrome). On the contrary, tools like Audacity, which are heavily based on C++ to optimize performance, are less suitable to easy system extension.

One last dimension of extensibility would be interoperability, with the aim of having one general purpose application to access and browse through data or even to do data analytic tasks on datasets. Here, again, web applications are well suited for interoperability as they are easier to maintain, update and develop than directly on the operating system installed tools. However, note that this requirement is going along with the performance requirement as one has to deal with the trade-off having a cross-platform application not specialized for one operating system and its lack of performance optimizations available in modern web browsers.

Configurability

At last, through the configurability requirement, we qualify the capacity of an annotation system to "offer a large panel of user-configurable functionalities dedicated to bioacoustic annotation". Raven Prox X is undoubtedly the most exhaustive annotation tool available today towards bioacoustics applications, while Audacity is the less fit-on-purpose.

Although priorities for this first version of APLOSE were more focus on accessibility and performance, we still propose basic campaign management tools like setting up an annotation campaign. We also allow the user to switch between different pre-configured spectrogram resolutions. Contrasting with other softwares, APLOSE is less flexible in the choice of spectrogram resolutions during a campaign, that needs to be pre-set at the campaign creation. However, its resulting advantage is to be highly performing regardless these resolutions and to propose a smooth resolution switching, while in other softwares each demand of a different resolution will ask for the re-computation of the displayed spectrogram, each time occurring a waiting time that can be unreasonably high when dealing with large volume of annotation data.

One original APLOSE feature is to be able to freeze annotation parameters during a campaign, which we found very interesting e.g. to perform parameter-specific sensitivity study. This is done via configurable permissions and access rights which are provided by the data owner to other Aplose users, and to our knowledge is not available in existing tools. Another original feature of APLOSE is the capacity of following the progress of an annotation campaign.

[17] `https://en.wikipedia.org/wiki/Extensibility`

Tools	**Features / Requirements**				
-	Accessibility	Scalability	Performance	Extensibility	Configurability
Raven Pro	X	X	X	X	√
Audacity	X	X	X	X	X
Koe	√	√	X	√	X

Table 1.1: Schematic comparison between desired APLOSE non-functional requirements and related existing features from three different annotation tools of bioacoustics community.

Chapter 2

System overview

2.1 On the user side

In this section we propose to the reader a short tour of APLOSE functionalities on the user side.

2.1.1 Preparing a campaign

User profiles We define two types of user profiles based on their role during the annotation process: the **administrator** (or campaign leader) and the **annotator**. The former is the person who will create the annotation project, upload the data, define the annotation tasks/ontology (i.e. which calls or acoustic event will be annotated) and the campaign parameters (e.g. spectrogram parameters, number of zoom levels, name of the labels...), as well as whether annotators can modify these parameters or not.

The annotators are the persons who use the interface only to annotate the data. Most commonly, they are invited to participate to a campaign with login details and the APLOSE url. They log in with a specific user name and password and then have access to the list of tasks they have to annotate, as illustrated in figure 2.1.

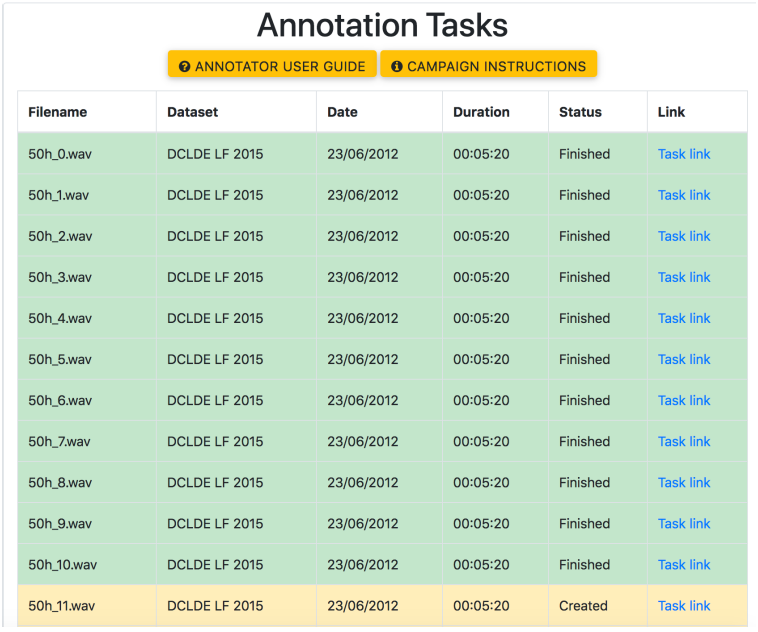

Figure 2.1: Annotation tasks window. The files that have already been annotated appear in green, the other in yellow.

Preparatory data In order to launch an annotation campaign, 3 CSV files are required. These files contain metadata about the the recording campaign and audio data such as recording sites, campaign name, campaign period, sampling rate, duty cycle and others. An exhaustive list of metadata to provide and templates of these files are available on github `https://.....`. Each annotation campaign has a unique name and a specific annotation task with predefined labels. A list of predefined annotators is also created. At the time of writing, the number of annotators cannot be modified once the campaign is launched. Moreover, the following parameters have to be pre-defined before launching the campaign :

1. largest duration of a displayed spectrogram ;
2. number of zoom levels, which will define the smallest duration of a displayed spectrogram ;
3. spectrogram configurations with different time and frequency resolutions.

One needs to provide the full set of spectograms to be displayed during the campaign and associated wav files. Note that only WAV audio files are supported for now. Further computational details on spectrogram pre-computing will be provided in Sec. 2.2.2.

2.1.2 During the campaign

Interface description As shown in Figure 2.2, the main page of the site allows to annotate a spectrogram visualization of a sound file. The spectrogram is labelled with time and frequency axes. Several control panels and buttons are available for the annotators to improve their user experience.

1. allows to change the way the spectrogram is displayed. With the select list, the user can choose the way the spectrogram was generated among available settings (nfft, winsize, overlap). A zooming feature is available by two means:

   - Clicking / tapping the buttons on the control panel: the spectrogram is centered on the progress bar
   - Scrolling over the spectrogram with a mouse or a touch pad: the spectrogram is center on the cursor position

   The spectrogram only zooms on time (no zoom on frequency). The zoom is discrete: each zoom level offers pre-computed spectrograms, meaning levels are decided by the creator of the dataset.

2. The sound file is playable by clicking on the play / pause button under the spectrogram. A thin black playback bar is displayed over the spectrogram (not shown on the Figure 2.2). Moreover, the speed at which the sound file is played can be changed from a select list displayed (only on Firefox) next to the play button. Available speeds are 0.25x, 0.5x, 1x, 1.5x, 2x, 3x and 4x. There is no pitch correction so modifying the playback rate also modifies the frequency. Listening to low frequency sounds is allowed thanks to this specific feature.

3. The **Submit & load next recording** button works this way:

   - If several annotations are not tagged, it selects the first one, display an error message and stay on this task
   - If all annotations are tagged (or if no annotation has been created), it saves them for this task, and loads the next available task
   - If there is no next available task, the user is sent back to the task list for this campaign.

   If this task has been annotated and submitted previously, the application will load and display previous annotations. Like new ones, these annotations can be modified or deleted.

4. To create an annotation, click on the spectrogram and drag over the area containing the feature. On click / tap release, the annotation is created and selected: it appears in **Selected annotation block** (4) and in the **Annotations list** (6), both below the spectrogram. Overlap annotations are permitted.

Moreover, annotated event can be listened to (at speed rate set in (2)) by clicking the play button at the upper left of the time-frequency box. It also can be easily removed with the close button at the upper right of the annotation box. A selected annotation block gives precise details about the annotation: start and end time, min and max frequency (4). It also list available tags (from the dataset). To tag / untag the annotation, press the matching button in (4) among the different choices set by the campaign manager (here "Dcall", "40-Hz", "Unknown call"). An annotation must have one and only one tag.

5. All the annotations created by the user for the current task are listed in the annotations list block, sorted by start time. Start / end times, min / max frequencies and the tag are ordered in row in this table. Clicking on an annotation selects it (it appears it the selected annotation block and can be tagged).

6. are links to the user manual for the annotation tool and the campaign instructions given by the campaign manager.

The red button **Back to campaign** at the upper right of the interface leaves the current task and saves the work in progress.

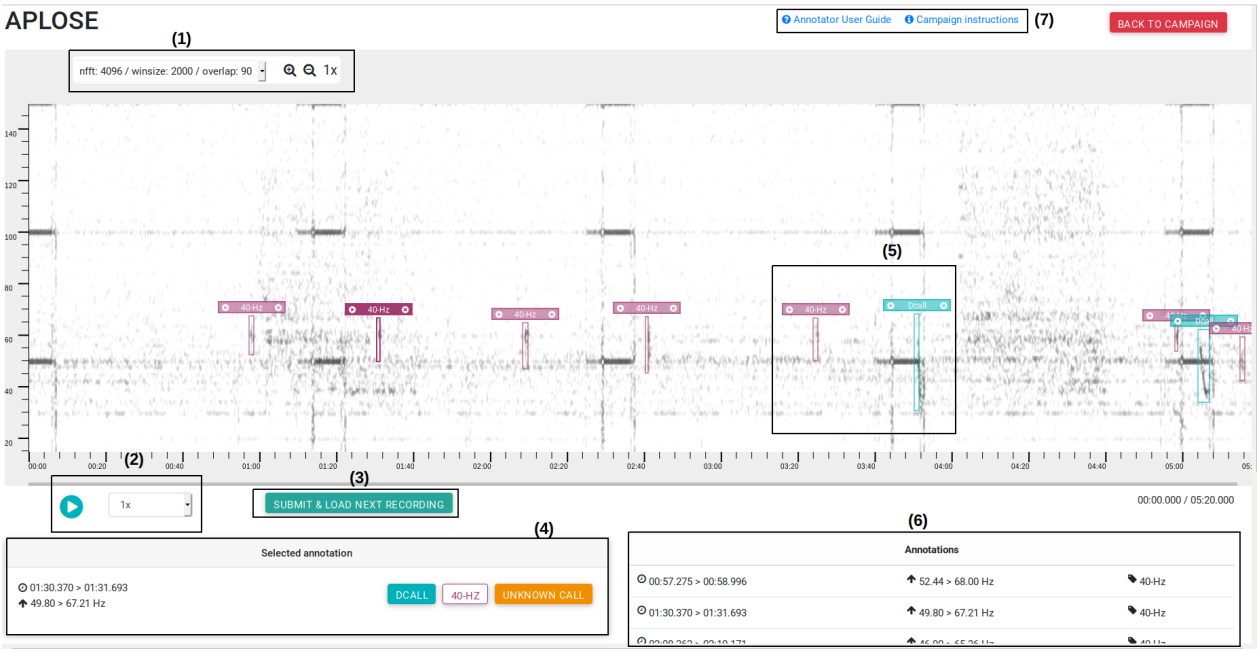

Figure 2.2: Layout of the APLOSE tool

Progress monitoring In the *Annotation Campaign* tab, by clicking on the campaign's name, a dashboard is displayed. It describes the campaign and details the progress of each annotator on this campaign. Administrators have access to a dashboard from where they can perform all these tasks. They can also annotate the data from their account as well as monitor the progress of the annotators and download the annotations as csv files. Annotators can check their own progression within each dataset, and modify or add annotations to the spectrograms. Once every audio file is annotated, the annotator can always go back and edit its annotations until the campaign is ended by the administrator.

2.1.3 Ending a campaign

The *Download CSV results* button creates a CSV file of all the annotations for this campaign. Each line of the file is an annotation, with the following columns:

- **dataset**: dataset short name

- **filename**: task file name
- **start_time**: start time of the annotation box
- **end_time**: end time of the annotation box
- **start_frequency**: min frequency of the annotation box
- **end_frequency**: max frequency of the annotation box
- **annotation**: the tag for the annotation
- **annotator**: the user who did this annotation

2.2 On the development side

2.2.1 Key components

APLOSE is an open-source, web-based tool programmed in JavaScript with React (for the front application) and Node.js (for the feature data API) libraries. It has been dockerized for an easy deployment on servers. The front-end part is heavily inspired by the extended version of wavesurfer used in the CrowdCurio project. Mozilla Firefox is full-featured whereas Google Chrome does not support the sound playback function yet. APLOSE source codes can be found on Github at `https://github.com/Project-ODE/` with a GNU GPL v3.0 license. The repository contains the source code, documentation about the installation and a user guide for annotation task. At the time of writing, the APLOSE system is hosted temporarily on server infrastructure within the OVH private company, with ongoing negotiations to transfer it permanently to the IFREMER server infrastructure named DATARMOR.

2.2.2 Spectrogram generation

Our spectrogram generation method lays on the distributed computing framework detailed in OSmOSE (2019), which include both standardized signal processing definitions and a distributed computing architecture that outperforms classical computing systems in processing high volume at scale with speed. A stand-alone Python custom code has also been written for local generation of spectrograms, available on GitHub `https://github.com/ixio/ODE-Scripts`. This code performs classical audio pre-processing operations such as filtering, amplification to raw audio signals to generate appropriate spectrograms, after pre-segmenting original files into smaller files with the desired duration as spectrogram displayed duration.

Note that current version of APLOSE does not allow to apply hydrophone or system calibration, and spectra are displayed in relative dB scaled to give a 0 dB maximum, such that amplitudes have negative dB values. Regarding this question, one interesting feature of the APLOSE version connected to a distributed back-end like OSmOSE (2019) is to quickly search for a more relevant maximum value (e.g. in a pre-defined biologically-relevant frequency band) with which normalizing the entire audio dataset, and set nominal contrast values around this value before starting a campaign.

2.2.3 Tile-based rasterization

Tiled server side rasterization on the other hand does not have any size limitations beyond your server farms capacity. And even that is not a hard limit as you can either pre-generate all tiles, spending whatever time necessary to do so, or generate tiles and cache them on the fly depending on your specific data and usage patterns. Tile-based rendering solutions have been initially developed for geospatial map applications, allowing to pan and zoom over the whole world. Loading one huge world image and just zoom and pan on it would lead to either bad image qualities on higher zoom levels or allocating too much memory and taking huge amounts of transmission time when using high resolution images. To avoid both problems the world image is split into tiles and stored in a tile tree structure, which typically starts with zoom level zero, which includes every geometry; zoom level 1 has four tiles and every higher zoom level doubles the amount of tiles.

Our requirement of handling large acoustic datasets naturally motivated this solution of pre-rendered raster images in our work. Similarly, the initial acoustic recording (zoom level 0) is sliced into tiles according to the number of zoom levels requested, as illustrated in figure 2.3. Individual spectrograms are computed for all tiles and then concatenated to provide a whole audio spectrogram at each zoom level. Our tile tree structure is a simple directory tree and PNG files on a hard disk.

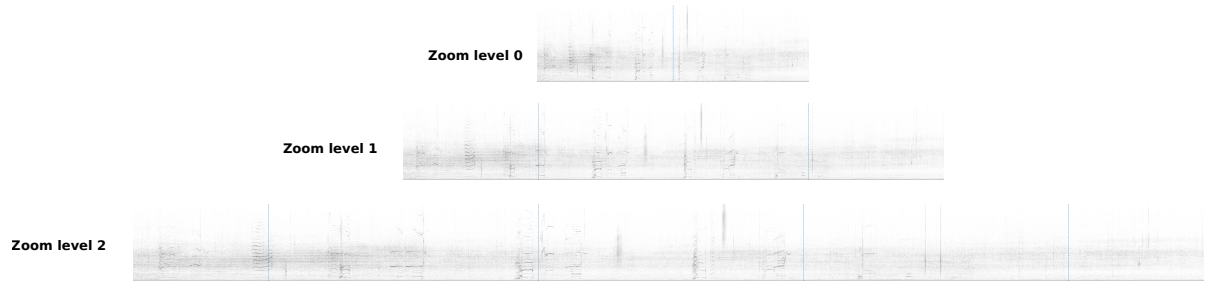

Figure 2.3: Illustration of the different zoom levels in the tile tree.

Chapter 3

Experimental evaluation and use cases

In this section, we first provide short experimental evidence on the performance feature/requirement mentioned in table 1.1 (Sec. 3.1), then we present the seed demo where anyone can test APLOSE on different datasets (Sec. 3.3), and finally we present in broad lines an ongoing research study case about inter-annotator variability using APLOSE (Sec. 3.2).

3.1 Experimental evidence for APLOSE performance

As already mentioned, APLOSE was built with the essential requirement of "offering a good-user experience, even if that experience shall involve high-volume datasets". In the following, we describe a short user-experiment demonstrating that this requirement is better met in APLOSE than in existing softwares, using Raven Pro X as comparative software.

We conducted two annotation campaigns with the DCLDE2015HF dataset (see presentation in Sec. 3.3) using parameters described in table 3.1. The first set of parameters (i.e. max display duration of 300 s, nfft = winsize = 4096, overlap = 90 %) corresponds to a "event-wise annotation setup at fine temporal scale", while the second one (i.e. max display duration of 1080 s, nfft = winsize = 1024, averaging_factor = x 10) corresponds to a "scene-wise annotation setup for long-term". With its high sample frequency of 200 kHZ, the DCLDE2015HF dataset is a good example of high-volume datasets in passive acoustics, where each 5 minute long audio file weights 114 MB.

Experiments were performed using a personal computer MacBookPro 2.7 Ghz i7 16 Go RAM. We must be very clear that these results do not intend to report reliable quantitative metrics but rather rough magnitude of order on waiting times that are significantly perceived by a software user. The measured loading times in Raven Pro X are around 20 s and 6 s per annotation window, respectively for the two campaign setups. Figure 3.1 provides more loading times as a function of the displayed spectrogram duration. Naturally, loading time of the spectrogram increases both with the display duration and the overlap ratio, impacting directly the number of Fast Fourier Transform to be performed, which is the operation with the highest computational cost in a spectrogram display. And of course, the higher the sampling frequency, the longer it takes to compute and load the spectrogram for a given display duration. Comparatively, APLOSE is by construction insensitive to these parameters, which only impact the creation phase of a campaign but not its progress. For these two sets of parameters, APLOSE exhibits waiting times inferior to 1 s between each annotation window, depending mainly on the user internet bandwidth.

3.2 Research study cases

In this section, we describe an ongoing research study aiming to better quantify and understand inter-annotator variability within collaborative annotation campaigns, so as to illustrate the potential of APLOSE of helping researchers to answer more fundamental questions in marine bioacoustics. In the following we briefly describe the annotation campaign setup and illustrative achieved results, and we also provide details on how joining this campaign and reproducing current analytical results.

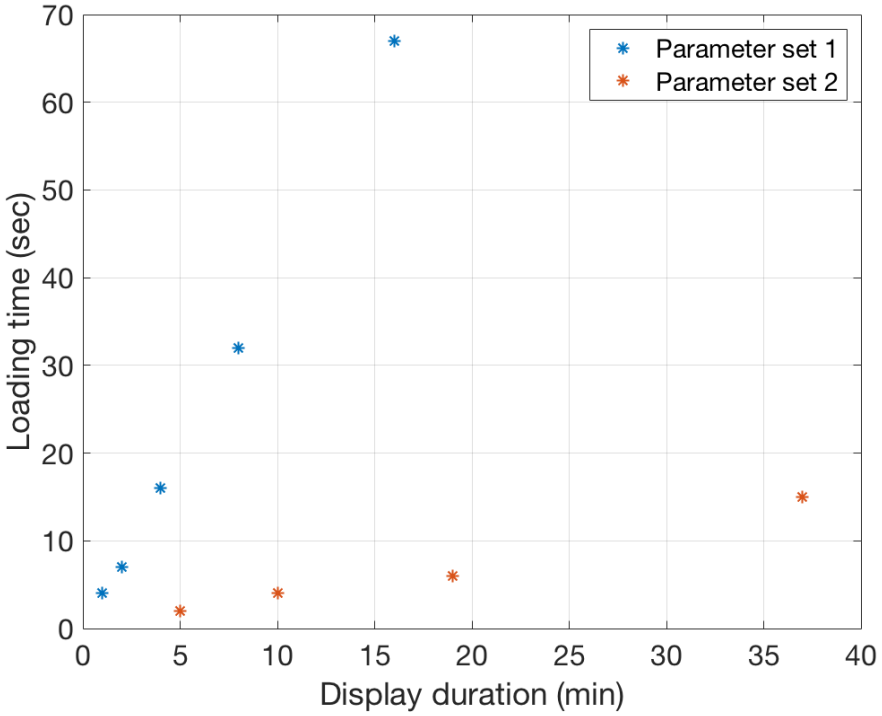

Figure 3.1: Loading times in Raven Pro X using two annotation campaign setups on the DCLDE2015HF as described in table 3.1.

3.2.1 Annotation campaign setup

Dataset The publicly available DCLDE2015LF dataset has been used (see presentation in Sec. 3.3). A 50h-long audio sampled at 2 kHz was splitted into 5min20s long smaller files. The dataset contained 563 small files to annotate. Annotators were asked to identify: D-calls (ref) and 40-Hz pulses (ref). If they have doubt about one of the two calls to find they could use a "Unknown call" label.

Campaign preparation Several parameters were set initially by an expert committee. The contrast and resolution of the displayed spectrograms, audio files were high-pass filtered at 15 Hz and low-pass filtered at 150 Hz, and then pre-amplified with a 35 dB gain, spectrograms were normalized based on the mean of maximum values of PSD from all wav.

Annotating process Annotators were given instructions (see Supplementary Materials) where visual but also aural examples of the sound to annotate were shown. They could refer to it at any time during the annotation process. Annotators could use a zoom from x1 to x8 on the spectrogram, speed up or decrease sounds. Annotations could only be made by visualizing and listening to spectrograms. Waveforms were not displayed for this annotation process.

3.2.2 Some achieved results

Assessing the inter-annotator agreement plays a key role to build more reliable and valid annotated datasets in the underwater passive acoustics community. Here, validity can be defined as the correctness of annotations (i.e "ground truth" or "gold label"). However, in underwater acoustics, there is no ground truth as annotations rely on perception and interpretation of annotators. In order to approximate golden labels, reliability of annotations is measured. Reliability can be defined as a measure of how consistent an annotation is across. Annotations are reliable if their agreement is high. A high reliability is a prerequisite for validity.

Number of annotations per annotator This is the number of acoustic events identified by each annotator. Noisy annotators (randomly annotate data) can be identified if their number of annotations is very different to others (cf Fig. 3.3).

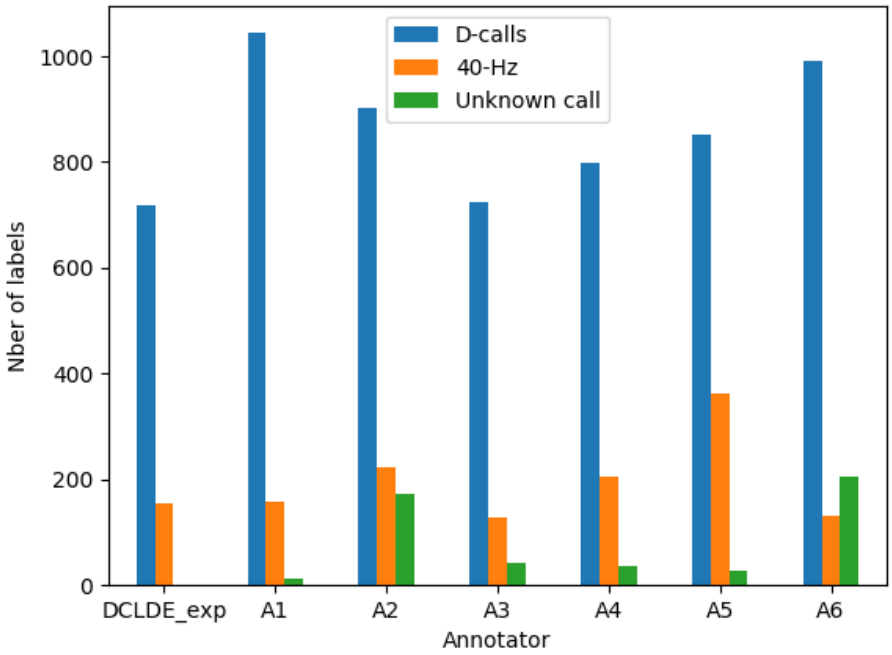

Figure 3.2: Number of annotations per annotator and labels.

Annotation task duration This is the time spent for annotating an audio file. At the end of an annotation campaign, a CSV file containing annotation times for each file per annotator is retrieved.

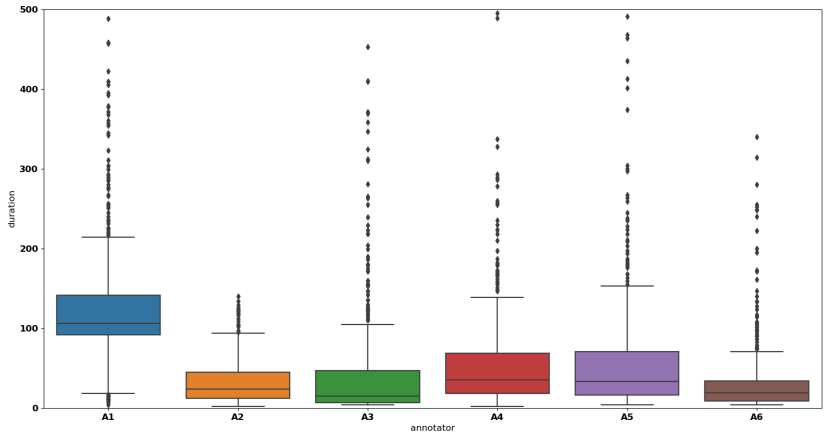

Figure 3.3: Boxplots on duration of an annotation task per annotator.

3.2.3 Joining and updating our DCLDE2015LF campaign

Leveraging the scaling out capacity of APLOSE (see Sec. 1.3.3), our DCLDE2015LF campaign can be joined by anybody, and annotation results automatically updated. To do so, please send a demand to the chief DCLDE2015LF campaign Paul Nguyen (p.nguyenhongduc@gmail.com), and he will give you access details. To our knowledge, these are the first attempts to provide to the marine bioacoustic community sustainable open collaborative annotation campaigns, which can be easily updated over a very long period of time by anybody. A first significant step towards crowd sourcing.

Furthermore, the csv result of any APLOSE campaigns is openly distributed directly on the web interface (even if you have not been enrolled as an annotator of the campaign). We alos distribute a jupyter notebook[1] that computes basic summary statistics on this csv file using python-based panda operators to make it easy for you to get started with such analysis.

3.3 Demonstration version of annotation campaigns

As a last evaluation step of APLOSE, and in complement to the DCLDE2015LF re-annotation campaign described above, we now provide access details to demonstration versions of ongoing annotation campaigns so that interested readers can directly experiment our tool. Three publicly available sub-datasets, described below, are used, exhibiting different sample frequency ranges and spectrogram parameter sets. These parameters are detailed in table 3.1, and have been chosen to cover different annotation use cases, e.g. from fine-scale event annotation to scene annotation over long-term averaged spectrograms. Access link to these demonstrations is `https://demo-aplose.osmose.xyz/` (login: dc@test.ode / pwd: password).

DCLDE2015LF The 2015 Low-Frequency DCLDE dataset (DCLDE2015LF) was recorded with High-frequency Acoustic Recordings Packages deployed off the southern and central coast of California at different locations, spanning all four seasons, over 2009-2013 period (see the full dataset documentation at `http://cetus.ucsd.edu/dclde/datasetDocumentation.html`). The three different recorders (CINMS site B, DCPP sites A and C) were resampled at 2000 Hz, and exhibit different depths of 600 , 65 and 1000 m, respectively. As this dataset was used in the DCLDE international challenge on detection and classification of marine mammals in 2015, it has already been annotated by two independent experts, with a total of 5211 strong labels (i.e. with start and end times of events) over 2 whale species classes that are highly unbalanced: blue whale D calls (4796 samples) and fin whale 40 Hz calls (415 samples).

OrcaSound Our second demonstration dataset is composed of recordings from the open source project Orcasound[2]. Centered within the summertime habitat of the endangered southern resident killer whales, Orcasound Lab is a good place to listen for orcas as well as ships passing through Haro Strait and boats traveling along the west side of San Juan Island.

We ingested in APLOSE one of their test set already annotated by Scott Veirs, that will be re-annotated by 5 other regional experts with the goal of reaching consensus on a label for each SRKW signal (calls only for now, not whistles or clicks). The test set is listed in the orcadata wiki under "Intermediate signal:noise ratio" and is 1/2 hour from 5th July 2019.

DCLDE2015HF The 2015 High-Frequency DCLDE dataset (DCLDE2015HF) was recorded at the same locations as the low-frequency one[3]. For this study, only the November 2009 SOCAL R campaign was analyzed. The mooring depth was 1200m. This dataset was also used in the DCLDE international challenge on detection and classification of marine mammals in 2015. We do not know which annotation guidelines and protocol was used to annotate it. During this campaign, acoustic encounters ("Any period of an animal echolocation that was separated from another one by five minutes or more was marked as a separate encounter. Whistle activity was not considered") from Sperm whale (5 acoustic encounters), Cuvier's beaked whale (2 acoustic encounters) and unidentified odontocete (1 acoustic encounter) were identified.

[1] `https://github.com/ixio/ODE-Scripts`
[2] `https://www.orcasound.net/portfolio/orcasound-lab-hydrophone/`
[3] See the full dataset documentation at `http://cetus.ucsd.edu/dclde/datasetDocumentation.html`.

Parameters	**DCLDE2015LF**	**orcasound**		**DCLDE2015HF**	
Sample frequency (kHz)	2	44.1		200	
Max → min display duration (s) / Zoom level number	**320** → 640/3	**60** → 3.75/4		**300** → 9.375/5	**1080** → 33.75/5
Original file volume (MB)	4.6	10.1		114	412
nfft (samples)	4096	4096	4096	4096	1024
winsize (samples)	2000	1024	4096	4096	1024
overlap (percent)	90	90	0	90	0
averaging factor	0	0	0	0	x 10
APLOSE tile volume (MB)	3.7	27.2	3.2	70.9	26

Table 3.1: Parameter description of the different APLOSE seed datasets. We also provide a comparison between the volumes in MB of original wav files and APLOSE annotation files, containing the pre-filtered audio files to be listened (wav files) to and the pre-computed spectrogram images png files) to be visualized. Note that the maximal spectrogram duration corresponds to the original wav file duration.

Chapter 4

Conclusions & Future directions

In this report, we have introduced the open-source web-based tool APLOSE whose main features are to allow for collaborative annotation campaigns on larger scales, both in terms of annotator numbers and data volume to be processed, than what can currently be done. Potential use cases are to help standardizing annotation processes and to build cross-validated reference datasets for the UPA community. We have demonstrated such needs with its evaluation by highlighting some disagreement between marine bioacoustics experts on a test case study.

We have not mentioned yet that such challenging goals were only made possible through a collaborative partnership between ocean researchers and software engineers, setting up locally at the Technopole Brest Iroise (French Brittany). In the future, we consider automating all preprocessing tasks before launching the annotation campaigns but also the inclusion of active learning methods to reduce the cost of the labelling budget.

Bibliography

Anorim, M.C.P. (**2006**). *Communication in Fishes* (Collin S.P., Moller P., Kapoor BG), chap. Diversity of sound production in fish, p. 71–105.

Bogaards, N.e.a. (**2008**). "Introducing asannotation: a tool for sound analysis and annotation." ICMC.

Bondi, A.B. (**2000**). "Characteristics of scalability and their impact on performance." In *Proceedings of the second international workshop on Software and performance – WOSP '00. p. 195. doi:10.1145/350391.350432. ISBN 158113195X.*

Boulton, G., Campbell, P., Collins, B., Elias, P., Hall, W., and Laurie, G.e.a. (**2012**). "Science as an open enterprise." Tech. rep., The Royal Society Science Policy Centre report 02/12. The Royal Society: London.

C. Cannam, C. Landone, M.S. and Bello., J.P. (**2006**). "The sonic visualizer: A visualization platform for semantic descriptors." In *Proceedings of the 7th International Conference on Music Information Retrieval (ISMIR-06), pages 324–327, 2006.*

Cagnacci, F. and Urbano, F. (**2008**). "Managing wildlife: a spatial information system for gps collars data." Environ Model Software. 23:957-9.

Cartwright, M.e.a. (**2019**). "Crowdsourcing multi-label audio annotation tasks with citizen scientists." ACM.

Fukuzawa, Y.e.a. (**2019**). "Koe: Web-based software to classify acoustic units and analyse sequence structure in animal vocalizations." Methods Ecol Evol. 2020;11:431–441.

Huettmann, F. (**2005**). "Databases and science-based management in the context of wildlife and habitat: toward a certified iso standard for objective decision-making for the global community by using the internet." J Wildl Manag. 69:466–72.

Kim, B. and Pardo, B. (**2017**). "I-sed: an interactive sound event detector." In *IUI '17: Proceedings of the 22nd International Conference on Intelligent User Interfaces, p. 553-557.*

Krijnders, D. and Andringa, T. (**2009**). "Soundscape annotation and environmental source recognition experiments in assen (nl)." In *Inter Noise.*

Leroy, E.e.a. (**2018**). "On the reliability of acoustic annotations and automatic detections of antarctic blue whale calls under different acoustic conditions." J. Acoust. Soc. Am.

Lowndes, J., Best, B., and Scarborough, C.e.a. (**2017**). "Our path to better science in less time using open data science tools." Nat Ecol Evol 1, 0160.

Marques, T.A.e.a. (**2013**). "Estimating animal population density using passive acoustics." Biol. Rev., **88**, 287–309.

Melendez-Catalan, D., Molina, E., and Gomez, E. (**2017**). "Bat: An open-source, web-based audio events annotation tool." In *Web Audio Conference WAC-2017, August 21–23, 2017, London, UK.*

OSmOSE (**2019**). "Theory-plus-code documentation of depam." Tech. rep., OSmOSE report 1, arXiv:1902.06659.

P. Herrera, J. Massaguer, P.C.F.G.M.K.N.W. and Fabra., U.P. (**2005**). "Mucosa: a music content semantic annotator." Proceedings of the 6th International Conference on Music Information Retrieval (ISMIR-05), 2005.

P. Wittenburg, H. Brugman, A.R.A.K. and Sloetjes., H. (**2006**). "Elan: A professional framework for multimodality research." In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC), pages 1556–1559, 2006.*

Piczak, K.J. (**2015**). "Esc: Dataset for environmental sound classification, 23rd acm international conference on multimedia, brisbane, australia, oct. 2015, pp. 1015–1018."

Richardson, W.J., Greene, C.R., Malme, C.I., and Thomson, D.H. (**1995**). *Marine Mammals and Noise* (Greeneridge Sciences Inc., Editor(s): W. John Richardson, Charles R. Greene, Charles I. Malme, Denis H. Thomson, , Academic Press), chap. ACOUSTIC CONCEPTS AND TERMINOLOGY, pp. 15–32.

Sjolander, K. and Beskow, J. (**2000**). "Wavesurfer - an open source speech tool." INTERSPEECH, volume 4, pages 464–467.

Tufto, J. and Cavallini, P. (**2005**). "Should wildlife biologists use free software?" Wildl Biol. 2005;11:67–76.

Versluis, M., Schmitz, B., von der Heydt, A., and Lohse, D. (**2006**). "How snapping shrimp snap: Through cavitating bubbles." American Association for the Advancement of Science, **289**, 2114–2117.