Digital Animal Sound Archive: a collaborative repository for bio-acoustics

Thomas Vandenberghe[1] (tvandenberghe@naturalsciences.be), Robin Brabant[1] (rbrabant@naturalsciences.be), Yves Laurent[1] (ylaurent@naturalsciences.be), Claire Brabant[2] (claire.brabant@natagora.be), Bob Vandendriessche[3] (bob.vandendriessche@rlhp.be), Wout Willems[3] (wout.willems@natuurpunt.be), Steven Degraer[1](sdegraer@naturalsciences.be), Ruth Lagring[1] (rlagring@naturalsciences.be)

[1]Royal Belgian Institute for Natural Sciences - RBINS (Belgium)

[2]Natagora (Belgium)

[3]Natuurpunt (Belgium)

A wide variety of animals produce acoustic signals or calls, that are in many cases species-specific. The use of these animal sounds in biological and ecological studies is widespread as they can be used to study species distribution, phenology, ecology and behaviour of organisms that are often visually elusive (e.g., marine mammals, bats). This results in extensive individual collections (tens of terabytes range) that are scattered in many different locations (e.g., scientific institutes, universities, environmental consultants, citizen scientists). A critical aspect of being able to learn from such large and varied acoustic datasets is providing consistent and transparent access that can enable the integration of various analysis efforts. Considering the data sizes, processes are hard to scale up. The overall objective of the Digital Animal Sound Archive (DASA) is to set up a robust data model, and a user-friendly web interface enabling Belgian bio-acoustic workers to collect, archive and explore biological acoustic data and accompanying metadata. The main partners in the project are RBINS and Natagora and Natuurpunt, two nature conservation and citizen science NGOs. Similar projects are ongoing abroad, and reaching out to these initiatives to share experience will be an integrated part of the DASA project. Therefore, specialists from the Muséum national d'Histoire naturelle (MNHN) in Paris and the British Trust for Ornithology (BTO) are part of the Follow-up committee.

The added value of this digital collection is manyfold: (1) to serve as a digital archive, (2) to add to the collections hosted by the Royal Belgian Institute of Natural Sciences (RBINS), (3) to serve as a reference collection of species and their behavioural acoustic calls, (4) to offer a validated dataset for the development of automated identification software tools and (5) to serve as a dataset for new ecological studies on dispersal of species and habitat preference.

As a proof of concept regarding the applicability of the data model and web interface the developments will first be focused on the recordings of bat sounds. The first release will be open to bat calls collected both on land and at sea. RBINS curates large offshore windmill nacelle bat monitoring datasets, which this project will finally expose.

Scope

Bio-acoustic data systems can be categorized by the way the origin of the sound is determined. These aspects can be combined with each other:

1. Sound collection aspect: determination by a submitter, before submission into an archive.
2. Citizen science aspect: determination by a website user (validator).
3. Incremental: determination in addition to that of the submitter.
4. Primary: first determination because the submitter didn't do it.
5. Reference sound catalogue: validated sound files are tagged as high-quality and offered as a separate category. These files can then be used in a sound library or as classifier training material.

6. Real-time classifier aspect: web service that, upon receipt of the sound file, makes a determination using a classifier (storage is optional). The determination may or may not be returned to the submitter.
7. Delayed classifier aspect: a (primary or incremental) determination by a classifier that is executed at a later time (asynchronously). This is done in bulk on a complete archive.

The DASA project will implement aspects 1, 2 and 3 and build a foundation to include the use of automatic classifiers. Multi-species group classifiers augment the value of audio recordings since they valorise ‘by-catch sounds’ such as crickets and birds. They are of greater value than most bat classifier software which tend to label all non-bat sounds as “noise”, whatever the origin. As a first step towards an analysis pipeline, the Tadarida-D and C software packages, together with a basic bat sonotype classifier [Roemer et al., 2021], have been dockerized, so that we can better distinguish between “noise” of biological, mechanical or acoustic origin. This leads to better storage space use, as “true noise” can be filtered beforehand. At this moment RBINS has reserved 50TB of local storage for the project, including backup. Long-term storage is foreseen on tape.

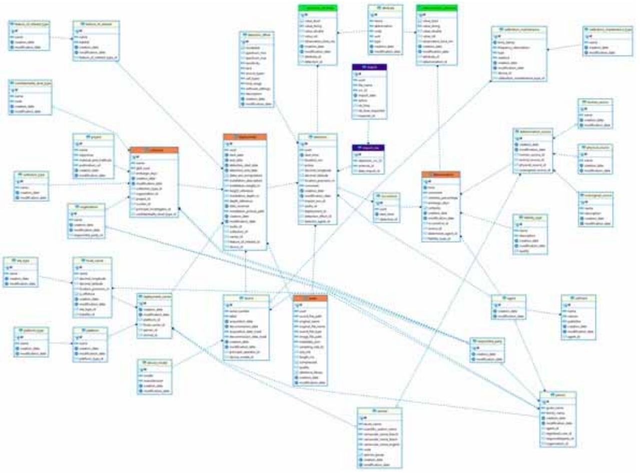

Figure 1 Data model.

Data model

The data model has been developed with fully normalised entities, to capture metadata to the fullest extent, for those who can provide it. After all, not all data providers have the same incentive for diligent metadata annotation. Many fields are optional to cater to volunteers and take the considerable workload needed for metadata annotation into account, especially for historical data and noise origin differentiation. The model is open to any species group and can describe long-term marine mammal surveys and sounds of physical or human origin as well (e.g., wind turbine piledriving at sea).

To develop the data model, we incorporated ideas from citizen science systems such as observations.org, Xeno-Canto.org and reference sound catalogs such as La Sonothèque and ChiroVox. Many aspects of the Tethys bio-acoustic data model have been taken over, such as the Deployment and DetectionEffort concepts. Tethys has been developed to manage the metadata from marine mammal detection and localization studies. The GBIF DwC Occurrence concept and the composition trough EventCore (deployment + detection) are easily retrievable from the data model.

Data acquisition

The project operates in three phases: 1) initial data acquisition from the partners in 2023, 2) web application development in 2024, and 3) a post-project operational phase where data is acquired via the web application (2025). In order to allow data transfer over http, the Tus.io protocol is strongly considered. To capture the metadata, a metadata Excel form has been created. The metadata form is more lenient than the data model and will be partially replaced by the web application. Finally, occurrence data will be disseminated to GBIF. Data owners can select a CC license; sensitive observations (commercial interests or rare species) can be embargoed at dataset or observation level.

Reference

Roemer C., Julien J.-F., & Bas Y., (2021). *An automatic classifier of bat sonotypes around the world*. Methods in Ecology and Evolution, 12, 2432 – 2444. https://doi.org/10.1111/2041-210X.13721