Universiteit Leiden

Master Computer Science

A FAIR Data Pipeline for Ecosystem Research, with Machine Learning and Marine Acoustic Data

Name: Jiamian He
Student ID: s3790525
Date: July 18, 2025

Specialisation: Data Science

1st supervisor: Mirjam van Reisen
2nd supervisor: Lu Cao
3rd supervisor: Burooj Ghani

Master's Thesis in Computer Science

Leiden Institute of Advanced Computer Science
Leiden University
Niels Bohrweg 1
2333 CA Leiden
The Netherlands

A FAIR Data Pipeline for Ecosystem Research, with Machine Learning and Marine Acoustic Data

Jiamian He

Abstract

The FAIR (Findable, Accessible, Interoperable, and Reusable) workflow offers a scientific paradigm for qualify analysis in big data era. In recent years, Passive Acoustic Monitoring (PAM) data have been collected for underwater ecosystem research purposes. However, these data are often dispersed across databases following different ontology standards. In this study, we review these federated marine acoustic data and ontologies, summarize the main soundscape-based machine learning algorithms for marine biodiversity assessment, and, through FAIR workflow engineering, propose an optimized data pipeline for FAIR ecosystem research. The findings provide a solution for researchers to integrate federated acoustic data for ecosystem research efficiently and to foster greater cooperation.

Keywords: FAIR Workflow, Data Pipeline, Marine Acoustic Data, Passive Acoustic Monitoring (PAM), Machine Learning, ontology

Acknowledgement

The work is supported by Leiden Institute of Advanced Computer Science (LIACS), Leiden University. I would like to express my heartfelt gratitude to Professor Mirjam van Reisen, who opened the door for me to understand the FAIR principles and inspired me with the beauty of data practices that connect disciplines, regions, and humanity. Through her guidance, I not only learned how to handle data responsibly, but also understood what it means to carry scientific spirit and social responsibility.

To my PhD guide, Joëlle Stocker thank you for coaching me far more than how to write a thesis. You showed me how to think scientifically, ask the right questions, and move forward step by step with patience and clarity. Your support and guidance shaped this research journey in every way.

I am also truly grateful to Dr.Lu Cao, Dr.Burooj Ghani, and Dr. John Graybeal for their generous feedback, encouragement, and sincere enthusiasm. Their scientific passion and kindness were a constant source of motivation throughout this project.

Table of Contents

1 Introduction

Passive Acoustic Monitoring (PAM) is a widely used method in marine research that uses underwater microphones (hydrophones) to continuously record ecological and biological sounds without disturbing marine life, providing accurate ecological data for studying marine ecosystems [1]. PAM allows researchers to monitor species presence, behavior, and population dynamics, as well as environmental conditions such as noise pollution or climate-related changes in the ocean. It significantly contributes to the assessment of biodiversity, detection of temporal cological shifts, playing a vital role in guiding conservation and management strategies [2]. Thus, the quantity of PAM data being collected has increased in recent years, with expanding spatial and temporal coverage.

With the rapid growth of PAM data from different sources, there is an urgent need to develop methods to integrate different data formats and process the data streams. The use of 'big data' approaches, including Artificial Intelligence (AI) and Machine Learning (ML), is important for solving problems in ecosystem science, as these methods can extract valuable patterns from large amounts of acoustic data. However, to apply these techniques effectively, metadata and data must be clearly structured, annotated, and accessible, enabling researchers to access catalogs, observations, and alert services via the webpage [3]. Data formats also need to support machine readability, interoperability, and scalability to align with the demands of scientific research and multi-source data inputs.

To meet the specific requirements of machine learning applications, it is also crucial to adopt suitable data frameworks that can handle preprocessing, feature extraction, analysis and annotation of the data in a consistent and automated way. These ML pipelines can quickly process large amounts of acoustic data to extract valuable information, assisting scientists to better monitor ecology and develop more effective conservation strategies. It is becoming a powerful tool and the main trend in ecosystem research. Some studies on ML architectures for PAM data exist, firstly on landscapes such as forests [24], and several studies have begun to apply this to marine acoustic data [3][17][24][25].

With the expansion of related research, there also has been a growing call for Open Science practice for better ecological study. In this context, the FAIR principles (Findable, Accessible, Interoperable, and Reusable) are essential. Some studies have

called for the FAIRification of marine data including image [4], metagenomic eDNA [5], and acoustic [6] data, which are widely used to monitor and explore ocean habitats. These data types are often massive and possess unique characteristics that traditional data management methods struggle to handle. Marine data are becoming increasingly diverse and inconsistent, while some countries have developed their own detailed data standards and metadata documents, these often differ significantly and lack universal adoption. Thus, effective data stewardship requires efforts across the entire data lifecycle - from collection and metadata documentation to quality control, publication, and archiving. To ensure long-term usability and cross-disciplinary collaboration, marine datasets must comply with FAIR guidelines, providing researchers with easy access and long-term use to answer new scientific questions [2].

Beyond the dataset, the computational workflow as digital object, is also important for Open Science. It outlines the multi-step processes in data collection, preparation, analysis, modeling, and simulation that produce new data products. The FAIR computational workflow [7] naturally aligns with FAIR principles by using existing metadata, generating new metadata during processing, and tracking data provenance. Such features improve the quality evaluation of ML data and support its reuse.

Ensuring Open Science and FAIR is not only a technical consideration but also essential for equitable access to marine acoustic data. In some regions, limited research capacity and unequal access to resources restrict joint scientific work. This imbalance can also cause gaps in marine acoustic data coverage, biased research results, and underrepresentation of local ecosystems in global studies [8]. Promoting FAIR and open science helps make these data and computational processes more accessible, encourages broader participation in marine acoustic research, and supports fairer contributions to policy and decision-making. Stronger international cooperation is needed to address these challenges and ensure ocean science benefits globally.

1.1 Problem Statement

Despite the advancements in PAM and ML applications, analyzing marine ecosystems with different ML algorithms remains challenging, due to varying input feature requirements and data processing methods. While some researches provide clear methods for data processing [3][24][25], they often overlook the specific needs of

downstream ML tasks. Also, although there has been progress in developing various data architectures and some studies applying ML to marine acoustic analysis [1][6][13][16][17][20][21], these frameworks mostly focus on single ML task. There remains a need to develop more adaptable frameworks that address diverse ML requirements for ecosystem research.

On the one hand, the raw acoustic data will be preserved separately by different organizations in different formats [2] (such as different databases and ontologies). There is a gap between standard ontologies and specific metadata definitions for machine learning tasks, such as the machine learning labels, etc.

On the other hand, creating pipeline that can effectively serve multiple ML tasks requires careful consideration, particularly how to structure it for various analytical needs including acoustic fingerprinting and anomaly detection still remains a challenge. Also, the complete computational pipelines are not always published, or are not always in a standard format with entire documents enabling reproducibility.

Therefore, for better computational practice and co-research, it is necessary to study how to bridge the gap between the requirements of data framework for multiple soundscape ML tasks, with corresponding ontology contributing to federated data interoperability and reusability, and the existing frameworks and ontologies. Given the complexity of marine ecosystems and increasing need for efficient data processing and AI analysis in marine ecosystem research, challenges such as data inconsistency, limited interoperability, and inefficiencies in handling diverse ML tasks persist.

Thus, **the goal of this research is to explore how to optimize data ontologies and pipelines to meet the multiple needs of soundscape ML tasks for marine ecosystem research, and to reflect on how to improve data interoperability and reusability.**

1.2 Research Gap

Three key gaps remain in the use of these techniques for ecosystem research:

- Incompatibility Between Data Ontology and ML Metadata: Acoustic data stored in a decentralized way lacks a unified metadata standard, hindering the interoperability and reusability of federated data.

- Insufficient standardized Multi-task Process Design: Existing frameworks cannot adapt to multiple ML algorithms, and soundscape level studies do not fully consider the needs of downstream individual studies. There is a lack of optimized and standardized workflows for multi-task processes.
- Insufficient FAIR computational workflow: Currently we focus more on FAIR datasets and algorithms, however the whole computational workflow should also be considered.

1.3 Research Objectives and sub objectives

To fill this research gap, the main objective is to propose an ML-driven data pipeline, by extended ontology, to enhance the reusability and interoperability of acoustic data for ecological research.

Specifically, the study focuses on the following objectives:

1) RO1: Integrate the marine soundscape data with optimized ontology to support subsequent ML tasks.
2) RO2: Address inefficiencies and propose the optimization solution of the data pipeline to support further diverse ML tasks.
3) RO3: In engineering, set up the pipeline and experiment to test its performance, and then document it.
4) RO4: Use the FAIR principle to evaluate the data pipeline maturity, as well as the data consistency and quality.

1.4 Research Question and sub-questions

Based on the research objectives above, the question is: How can the marine acoustic data pipeline and ontology be designed to address data consistency, support various ML tasks, and align with the FAIR principles for effective ecosystem research? Following this main question, the key sub-questions are:

1) RQ1: How can marine soundscape data be integrated, through ontology design to better support machine learning requirements?

2) RQ2: What are the key inefficiencies in current data pipelines to be optimized?

3) RQ3: In practice, how does the optimized pipeline perform driven by the data models?

4) RQ4: What degree of FAIR maturity does the data pipeline achieve?

1.5 Location

This research is conducted at the Leiden Institute of Advanced Computer Science (LIACS), Leiden University, as part of a master's thesis project. The study is literature-based and code-based, involving publicly available academic paper platforms, datasets and cloud-based resources. The only data collection involving human participants in interviews, which are conducted either offline in Leiden or online via Zoom. The research is carried out under the supervision of academic advisors at Leiden University and includes discussions and interviews with researchers from Naturalis Biodiversity Center(Leiden) and San Diego Supercomputing Center.

1.6 Relevance considerations

This study addresses an important challenge in marine ecological research: how to better use sound data with modern AI tools. The relevance is in two aspects:

- Academic Relevance: This study helps improve ecological research by showing how marine acoustic data can be used in a machine learning workflow that supports different tasks. It provides a clear way for downstream tasks such as species or noise source identification, where sound clips can be directly extracted from continuous soundscape data without reprocessing long recordings. This makes it easier to use the same data for both biodiversity studies and detailed species analysis. It also contributes to discussions on how to make data and workflows more reusable, consistent, and aligned with machine learning methods. As a case study, it shows how FAIR principles can be applied in ecological science.
- Societal Relevance: By making marine ecosystem monitoring more efficient, this study supports analysts, conservation groups, and policymakers. It helps with faster biodiversity assessments, early warnings of environmental change, and

better conservation decisions. The workflow can also help build scalable and open tools to protect ocean ecosystems in the face of climate change.

1.7 Philosophy of Knowledge

This study mainly takes a positivist position, through engineering to verify the optimized design. This study is situated in a specific case — marine acoustic data and FAIRification for marine ecosystem monitoring — using real-world datasets and domain-specific ML tasks, which makes the context highly relevant. However, the proposed workflow is transferable and can be applied to broader federated data analysis tasks.

1.8 Research Design

This study combines a case study and exploratory design to investigate how the FAIR principles can support the development of a machine learning (ML) workflow for marine acoustic data. The case study works with real-world PAM datasets, identifying key data analysis requirements for ecosystem research. It includes qualitative research, such as expert interviews and literature reviews, to understand the essential problems and use cases. Based on the challenges identified in the case study, the exploratory part focuses on integrating the data into a metadata-driven FAIR pipeline and exploring how well it can support different ML tasks. The study examines key aspects of FAIRification, such as data linking, aligning ML tasks with available data, and designing the overall workflow. It also identifies practical challenges, including scalability issues and the lack of standardized, reproducible workflows for marine acoustic data analysis.

1.9 Methods of data collection

This study collects data in two methods:

Method 1 - Qualitative research data: The qualitative data is collected through expert interviews and literature review. The interviews were conducted with researchers in the field of acoustics data, focusing on the integration of data into machine learning for biodiversity research. Additionally, a systematic literature review of relevant

academic papers and technical reports was carried out to summarize the existing frameworks and algorithms. Those documents are all from public databases (Google Scholar, ScienceDirect, ResearchRabbit).

Method 2 - PAM marine acoustic data and metadata files: The soundtrack are all from publicly accessible platforms, one is reference published papers, and the others are from some national ocean sound databases including those in the United States, German, and Australia. All the metadata files can be downloaded with the soundtrack data. The entities for ontology design are from ontology webpages and portals such as BiodivPortal. Here is the summary of the data sources as Table 1:

	Data Source	**Year**	**Location**	**Data URL**	**Metadata URL**
1	NOAA-ONMS	2023	US	https://console.cloud.google.com/storage/browser/noaa-passive-bioacoustic/onms/audio/fgb01/onms_fgb01_20230714/audio;tab=objects?pageState=(%22StorageObjectListTable%22:(%22f%22:%22%5B%5D%22))&inv=1&invt=AbtAZg&prefix=&forceOnObjectsSortingFiltering=false	https://storage.googleapis.com/noaa-passive-bioacoustic/onms/audio/fgb01/onms_fgb01_20230714/metadata/ONMS_FGB01_20230714.json
2	Williams et al paper [25]	2024	French Polynesia	https://zenodo.org/records/10539938	https://zenodo.org/records/10539938
3	PANGAEA	2020	German	https://opus.aq/portal/recorder/ARKF04-19_SV1088?timestamp=2020-01-09T01%3A41%3A05.170%2B00%3A00	https://doi.pangaea.de/10.1594/PANGAEA.967512
4	AODN	2018	Australia	https://catalogue-imos.aodn.org.au/geonetwork/srv/eng/catalog.search#/metadata/e850651b-d65d-495b-8182-5dde35919616	*the same as data url

Table 1: Summary of the PAM data sources

1.10 Methods of data analysis

For the literature review, the PRISMA method was used for systematic review. Expert interviews were analyzed by reviewing notes, extracting key insights, and summarizing core recommendations. For the acoustic data, data linkage with unified metadata files and machine learning algorithms were the main ways for data integration and analysis. More details will be explained in Chapter 3 - Methodology part.

1.11 Ethical and data management

All datasets used in this study are publicly available and used only for academic, non-commercial purposes. According to their terms of use and licenses, proper attribution and citation are provided for each source. Since the research does not involve any sensitive or personal data, it complies with relevant legal and ethical regulations regarding data use.

To ensure FAIR findability, accessibility and reusability, all original datasets are uploaded and opened via Zenodo (*https://zenodo.org/records/15185049*). All experiment process data and source code used in the computational workflow have been published on GitHub (*https://github.com/holeiden/fair_thesis*), and the workflow is standardized with Snakemake and published in WorkflowHub *(https://workflowhub.eu/workflows/1380)*.

1.12 How can the knowledge to be generated be used

This research produces knowledge that is both specific to a particular context and partly transferable. The specific context includes the datasets, ontologies, and machine learning algorithms specifically used for soundscape research. The transferable part is the standardized FAIRification workflow developed in the study. It follows four main steps: 1) integrate federated data into ontology-based metadata; 2) analyze the metadata requirements of multiple ML tasks as input features; 3) FAIRify the computational workflow; 4) perform a FAIR maturity assessment. This workflow could be referred to many other research areas that deal with non-sensitive data.

The reliability of the findings is supported by the fact that the workflow follows the FAIR principles strictly. All data and code are shared openly, and the process is standardized using Snakemake, which makes it portable and reproducible in other environments.

1.13 Timeline

The research lasts for 9 months, and the timeline is as follow Table 2:

Oct 2024	- Kickoff, discuss with supervisor and PhD guide, define the research scope. - Start literature review and finding the research gap, consider about the topic.
Dec 2024	- Finalize the topic and research scope, deliver a research proposal. - Plan the methodology for the whole experiment part.
Jan 2025	- Data source, ontology & ML algorithm study, use tools to design the ontology. - Start writing the Chapter 1 &2.
Feb 2025	- Data pipeline engineering on different soundscape ML models. - Start writing the Chapter 3.
March 2025	- FAIR maturity evaluation. - Start writing the whole thesis, mainly on result chapters.
April 2025	- First draft of thesis. - Poster presentation in Solid Symposium 2025.
July 2025	- Finalize the thesis, defense and prepare for publication

Table 2: Timeline of the research

2 Theoretical Framework

This chapter presents the theoretical concepts and frameworks essential for understanding the integration of metadata & ontology, machine learning, and the FAIR principles in the upcoming research. These components form the basis for developing efficient FAIR computational workflows, ensuring that data is structured, machine readable and actionable, and reusable for analysis.

2.1 Metadata and Ontology

Interoperability is the ability of data or tools from different, non-cooperating sources to integrate and work together with minimal effort. Metadata and ontology are key elements that enable this interoperability.

Metadata is '*data that provides information about other data*', it can help users find relevant information and discover resources (*Wikipedia*). In another word, *'Metadata are contextual data about your experimental data. Metadata are the who, what, when, where, and why of these data. Metadata put these data into context' (https://microbiomedata.org/introduction-to-metadata-and-ontologies/).*

To ensure interoperability across different platforms, metadata terms should follow standardized vocabularies from the Semantic Web schemas, these schemas allow metadata terms to be consistently described, interpreted, and shared across various applications and systems.

Triplestore: The metadata can be structured using a triplestore approach, which represents relationships using the format subject – predicate - object, the subject is the individual being described, the predicate defines a property or relationship, and the object gives the value or target of that relationship. Triplestore is particularly powerful for representing complex, linked information in a way that supports interoperability and reasoning, making them well-suited for semantic web and ontology-based applications.

Ontology: is '*a formal description of knowledge as a set of concepts within a domain and the relationships that hold between them. It ensures a common understanding of information and makes explicit domain assumptions thus allowing organizations to make better sense of their data'*

(*https://www.ontotext.com/knowledgehub/fundamentals/what-are-ontologies/*). Ontologies help in organizing and linking metadata in a meaningful way, supporting more complex reasoning and data integration. This structure allows for semantic querying and reasoning, including class hierarchies, constraints, and logical rules. The Resource Description Framework (RDF) and Web Ontology Language (OWL) are widely used formats to represent ontologies in this triple-based structure.

JSON-LD: JavaScript Object Notation for Linked Data (JSON-LD) provides a lightweight, web-friendly and human-readable version to contain the ontology information. It allows data to be semantically annotated and integrated with Linked Data principles, using a Dictionary as another format to represent the triplestore relationship.

It has seamless integration with web technologies, such as being directly embedded in HTML pages or APIs, and also supports incremental construction and partial updates of metadata. This means that metadata can be progressively enriched or adapted to evolving research needs as new data is accumulated. Such flexibility is especially valuable in machine learning workflows for ecological data, as it changes such as updating the Anomaly Event labels, supporting sustainable data management and reuse.

Semantic filters and data query: Metadata not only describes data but also serves as a semantic filter, helping users efficiently locate relevant datasets based on key attributes. It enhances the ability of machines to automatically find and use the data, in addition to supporting its reuse [9]. With data query languages like SPARQL, users can query and extract individual data records in a batch that meet specific criteria, enhancing efficiency in large-scale data selection.

2.2 ML algorithms for marine acoustic data

Institutions have gathered extensive ecosystem data over decades, with various sensors on various platforms. To handle the growing data volume, they now use big data analytics, including AI and machine learning, to support resource management, monitoring, and policy decisions. The missions of acoustic data collection and analysis have expanded from single-species identification & statistics to overall landscape biomass estimations.

As shown in Figure 1, ML tasks with acoustic data are generally in two directions. One focuses on the ecosystem, analyzing patterns at soundscape level to study biodiversity or detect unusual changes. On the soundscape level, researchers often apply unsupervised learning method to process data, using CNN-based feature extraction and clustering methods to assess fish biodiversity or coral reef biodiversity using large historical datasets. It also produces the sparse dot map called 'sound fingerprint', which shows the latent features of sounds across time and space.

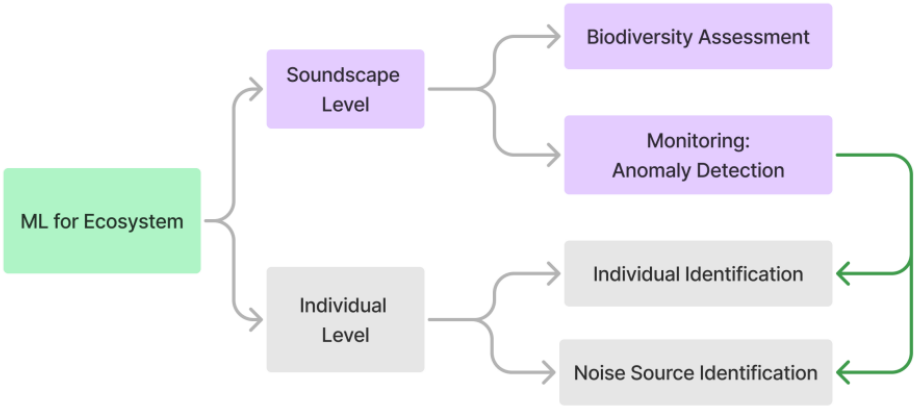

Figure 1: Different ML tasks of marine acoustic data for ecosystem

Another key application of soundscape data is anomaly detection—spotting unusual events in the acoustic environment. This is done using methods like RNNs, K-means, or Gaussian Mixture Model (GMM). Detecting anomalies early allows for faster response to disturbances, helps reduce harm to species, and improves the resilience of marine ecosystems. This is critical for effective biodiversity conservation.

The other focuses on individual-level analysis, such as identifying specific species or detecting noise sources, mostly are supervised learning methods. For individual-level tasks, many algorithms already exist to recognize particular species or classify different sound sources. However for individual-level machine learning, one common challenge is the sparsity of sound events (usually just 1-5% duration of the whole soundtrack). In long-term PAM monitoring, the actual sound events are often very short, and researchers usually need to extract these event windows from long recordings.

Fortunately, in soundscape-level analysis, the data is already resampled into standardized short sound clips. Some of these clips contain the events and can be

shareable and reusable for individual-level tasks, this can turn one unified dataset serving multiple different tasks efficiently.

2.3 FAIR principles

The FAIR principles aim to overcome barriers to data discovery and reuse by helping all stakeholders like researchers, institutions and systems, more easily find, access, integrate, reuse, and properly cite large volumes of data produced by modern science [9].

FAIR concept [9]: The FAIR Principles were firstly introduced by Wilkinson et al. (2016) to improve the infrastructure supporting the reuse of scholarly data. FAIR stands for Findable, Accessible, Interoperable, and Reusable, aiming to ensure that data and metadata can be easily shared and reused by both humans and machines.FAIR principles emphasize:

Findable: Data should be assigned a globally unique and persistent identifier (e.g., DOI) and be described with rich metadata indexed in searchable resources.

Accessible: Metadata and data should be retrievable using standardized communications protocols, even when the data itself is no longer available.

Interoperable: Data should use a formal, accessible, shared, and broadly applicable language for knowledge representation (such as RDF, OWL or JSON-LD, to enable integration with other datasets).

Reusable: Metadata and data should be richly described with clear usage licenses and detailed provenance to support replication and further use.

The FAIR principles are not a standard, but a set of high-level guidelines. They are widely endorsed in scientific domains, especially those dealing with complex and distributed data such as environmental and biomedical sciences.

Data FAIRification process: Data and metadata formats and processes are often very different between fields and organizations, which creates information silos and makes it hard to share data across scientific communities. The 'FAIRification' is a method that follows a step-by-step, generic workflow for making data FAIR. It can be implemented through collaborative workshops or by teams of domain experts working

under the guidance of FAIR data stewards, as outlined by Jacobsen et al. at Leiden University Medical Center and the GO FAIR International Support and Coordination Office [11].

Federated Architecture: Federated architecture is a solution to manage data in different databases but can still be accessed and used together without loading to a centralized storage. After data FAIRification, users can quickly link and query data across different institutions and storage systems. This decentralized approach respects data ownership, supports different types of data structures, and helps share and reuse data smoothly across organizations and regions.

FAIR computational workflow: is designed to manage complex, multi-step data analysis processes. They connect different pieces of code, software, and tools into a structured pipeline, and automatically manage how data flows between them. They are important because they improve efficiency, make research easier to reproduce, scale up better for large datasets, and support teamwork. With the rise of data-driven science, well-designed and FAIRified workflows help ensure quality, transparency, and the smart use of computing resources. Metadata plays a key role in describing what each step needs to run properly [12].

FAIR maturity: often referred to as 'FAIRness', represents the extent to which data objects, metadata, and associated workflows conform to the FAIR principles, that they are Findable, Accessible, Interoperable, and Reusable. FAIR maturity is commonly assessed using a set of structured indicators or levels, each corresponding to a specific sub-principle as Figure 2 below. These levels are derived from formal interpretations of the FAIR principles and provide concrete, measurable criteria for evaluation. By assessing FAIRness, institutions and data stewards can monitor the progress of their FAIRification efforts, benchmark against community expectations, and support interoperability and reusability across disciplines, platforms, and repositories. In this context, FAIR maturity model plays a role: they guide the implementation of FAIR-aligned practices and provide the metrics necessary for transparent, reproducible, and trustworthy data stewardship.

Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

Figure 2: FAIR Guiding Principles [9]

2.4 Relationship between Ontology, Machine Learning, and FAIR Principles

For the single-source data or an individual researcher, FAIR helps to make the dataset reusable for open science. Also, in a federated architecture, FAIR will be helpful to guide the integration process. When gathering different data source for marine biodiversity analysis, raw acoustic data needs to be processed into FAIR data so it can be reused effectively. This is especially important because data often comes from different institutions, which may use different recording equipment, file formats, or store data in different ways. Without standardization, it cannot be the inputs into the machine learning (ML) pipeline properly.

FAIR helps bridge the gap between diverse federated data and the unified input requirements of ML. It ensures that datasets are resampled and combined with clear and consistent metadata, structured through an ontology. Ontologies provide a top-down structure with shared vocabulary from semantic web, defining what each data element means, what is the unit, how it was collected, and how it connects to other datasets. This is essential to prepare data for ML workflows.

As shown in Figure 3, the core logic of the framework is to take raw acoustic data, process it through a data calibration and standardized process, and link the data samples using ontology-based metadata. After this it is possible to resample data,

extract standardized features from different sources, enabling ML models to analyze the data more effectively, compare results, and adapt to various tasks.

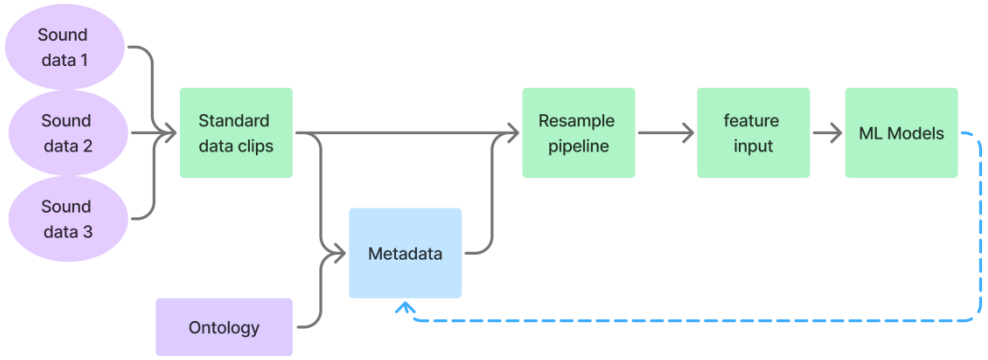

Figure 3: The framework of ML tasks for sound data in various ontologies

To support supervised training and preserve the results of unsupervised learning in acoustic-based ML tasks, metadata plays a crucial role in structuring and labeling datasets. In supervised learning (at the individual level), models rely on labeled data — such as species names, behavior types, or noise categories for training. Therefore, clearly defined metadata fields are needed to standardize labels (e.g., using scientific species vocabularies).

In contrast, unsupervised learning does not require labeled inputs, but updating metadata with anomaly detection results can provide a fast retrieval mechanism for subsequent individual-level analysis. From both perspectives, embedding machine learning labels directly within metadata ensures interoperability and reusability.

In short, FAIR principles guide the method to integrate federated data, extend metadata/ontology to form the foundation for interoperable inputs and outputs for machine learning. After FAIRification, each individual data is represented in a standardized, machine readable format. For the federated architecture (a decentralized data infrastructure), such FAIR data can be stored in its original location while being accessed and queried via standardized protocols. This enables data sharing and reuse across institutions and locations without the need to physically centralize the data.

3 Methodology

The FAIR principles provide a paradigm for managing data, extending the ontology, and ensuring data interoperability and reusability. This study start from data integration and the development of our ontology, followed by a review to identify various algorithmic applications. Then we design and implement the data framework, and finally evaluate the FAIR maturity of the data pipeline.

Before starting the actual data integration, we need to first study the metadata information. The data sources contain two types of metadata, one is general metadata as ground true information, such as sampling depth. These can be integrated by simply unifying the terms and units. The another one relates to acoustic differences because of the background noises or variations in the hydrophones used across different projects. These bias are stem from the instance features of the noise or the hydrophones themselves, rather than from actual features in the soundscape.

Thus when dealing with the second type of metadata, it is important to include relevant acoustic-related metadata during the FAIRification process at the start in Section 3.1, as this will support later data calibration. This remains information such as the original recording equipment, sampling frequency, and calibration details, providing important contexts for data calibration before data processing and ML in Section 3.2 and Section 3.3 to adjust for variations introduced by different recording devices.

3.1 Data source and ontology

This research begins with identifying suitable data sources and ontologies as the first step of the data FAIRification process (Figure 4). The Data FAIRification process shows a detail guideline to address these challenges, providing a step-by-step approach to identify and analyze, then define the data and metadata and link them both together, finally public and assess it.

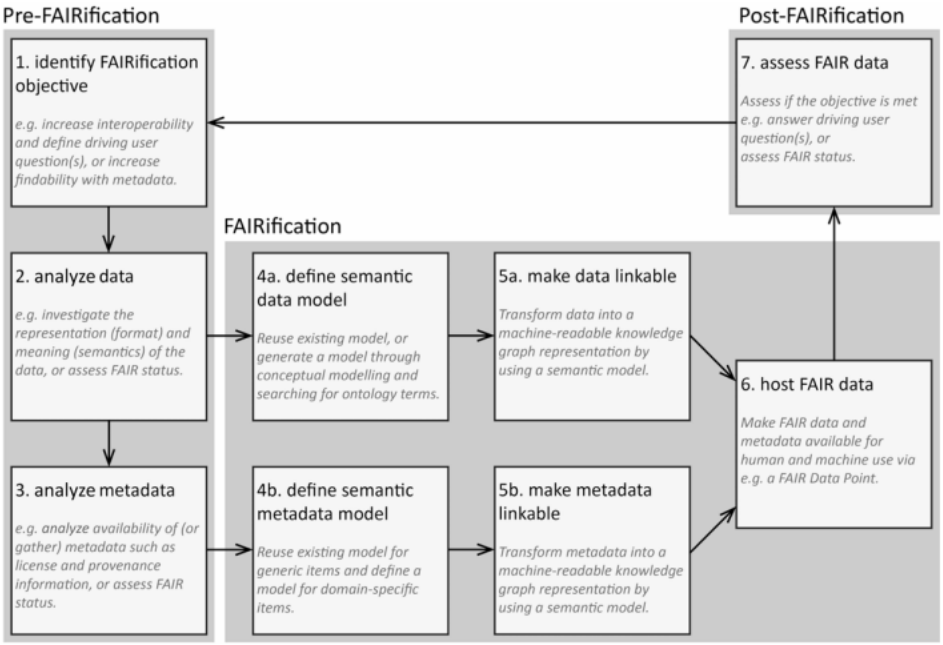

Figure 4: Standard data FAIRification process [11]

Figure 4 illustrates the FAIRification process, structured into three phases: Pre-FAIRification, FAIRification, and Post-FAIRification. The Pre-FAIRification phase involves defining the FAIRification objective (e.g., improving interoperability or metadata quality), and analyzing the current data and metadata, including their formats, semantics, and FAIR status. In the FAIRification phase, semantic models for both data and metadata are defined or reused, followed by transforming them into machine-readable, linkable knowledge graph representations. These are then hosted on platforms to ensure accessibility. Finally, in the Post-FAIRification phase, the FAIRness of the data is assessed to see whether the initial objectives have been achieved. This 7 steps' order is not strict and can be iterative.

In the Pre-FAIRification step, we focus on 'Marine Acoustic Data' as the core object to be integrated. To find widely adopted marine acoustic data sources and ontologies, we used three main methods: exploratory search using Google with keywords such as 'marine acoustic data' and 'marine PAM data'; review of academic literature to identify what datasets and ontologies have been used in previous studies; and interviews with acoustic and ontology experts, asking them to recommend data platforms they use and ontologies they follow. These expert inputs help validate our search and fill in the gaps that may have been missed.

We conduct interviews with researchers in two areas: data sources and ontologies. The interviews were carried out both in-person and online, each lasting around one hour.

Before each interview, we prepare a structured list of questions to guide the discussion. The responses are recorded through written notes, from which key information is later extracted.

In the FAIRification and Post-FAIRification step, after identifying several main marine acoustic datasets and ontologies, we select three representative datasets, each base on a different ontology structure. We then carry out these four steps of FAIRification:

1) Define a semantic metadata, and use Protégé to define the ontology model. Reuse terms from existing semantic web vocabularies as much as possible, such as Schema.org (for general-purpose metadata) or domain-specific schemas like Dublin Core. When defining metadata, terms (such as Duration, Locations, Event, Frequency) are identified and linked using Uniform Resource Identifiers (URIs). These URIs provide globally unique references, enabling different datasets to refer to the same concept and thereby facilitating linked data. Ontology portal platforms can be used to efficiently search for existing terms and classes. If required terms are missing, we create our new terms as extension. Once the model is defined, output the ontology as a Turtle file (TTL).

2) Check feature coverage: Prepare a checklist of important features and filters needed for ML tasks (e.g., location, duration, species label, frequency, sensor type). Verify whether these features and filters are already represented in metadata. If gaps exist, add ontology extensions to include the missing ones. The research on the FISHGLOB dataset [10] as shown in Figure 5, shows an example to consider the filters, which integrates fish biodiversity data from scientific bottom-trawl surveys, three types of filters can be used as selectors in data platforms. These filters allow users to navigate the data based on different aspects of the dataset, such as location, time, or species. The granularity of the filter directly enhances the precision of data selection, enabling more specific and tailored queries based on dimensions such as spatial and temporal scale, or taxonomic depth. By using these filters, users can easily find the data they need, making the dataset more accessible and useful for research.

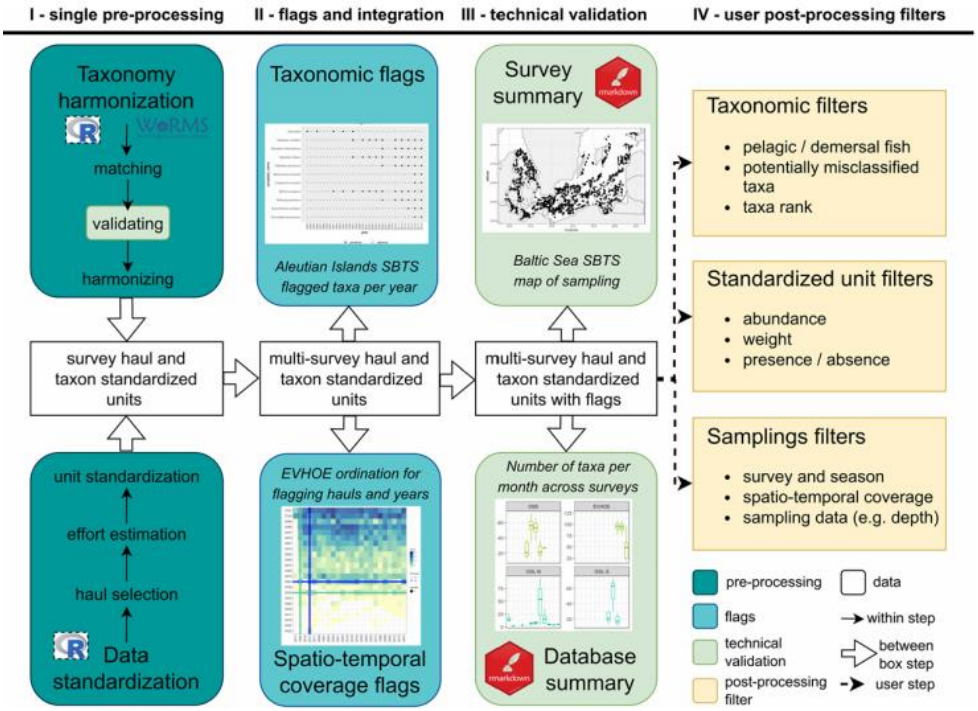

Figure 5: Example to create FISHGLOB datasets and three types of filters [10]

It is important to note that some metadata such as result labels generated by machine learning models or anomaly event descriptions are annotated as assertions or statements. These labels different from the original background metadata of the data as provenance features, so they are not ground truth and should be clearly distinguished. From a data lifecycle perspective, before downstream machine learning tasks, these labels can be further refined through expert review or cross-validation with multiple data sources and evidences. This helps reduce the risk of downstream task failure caused by the inaccuracy of ML-updated labels or manual event descriptions. These consideration should also be adopted in the Section 3.2 data framework design and the Section 3.3 practical pipeline engineering. FAIR is not static, by clarifying which labels are available for re-evaluation, the robustness of the entire data framework can be improved.

3) Data & metadata linkage: After calibrating and standardizing the raw long acoustic data, link every standard data as individuals with the semantic metadata. Because the original metadata files contain a lot of background information such as location, time, equipment and signal information etc, but probably with different representations, we need to make mapping rules to preserve the relevant metadata values. The mapping rules should consider three aspects: term standardized (e.g.

deployment location & project location all transferred to 'Location'); semantic standardized ('Depth' means hydrophone's deployment depth not the seabed's depth); unit standardized (Duration should always be in seconds).

Once we have the mapping rules, we could transfer individuals' original information into a unified table with standardized metadata. For instance, in the table format as Table 3, the value could be transfer into triples as <row><term (column)><grid value>: e.g. <id_1><Event>'1', <id_2><Location>'Stetson Bank'. Here each row could be transformed into a series of triples, this structure provides semantic clarity, allows for data linking across datasets, and supports flexible querying in triplestores.

id	Duration	Location	Event
1	5	Random	1
2	60	Stetson Bank	0

Table 3: Example of a unified table

Then we use a Python script finish the data & metadata linkage, transferring the metadata into JSON-LD format, combing the structured ontology information in TTL file and the individuals' metadata values in the unified table. Now with the metadata as filter, data queries can be implemented. SPARQL is a powerful query language designed to work with RDF files like OWL and TTL data formats, and it allows for the querying of complex relationships and inference rules within the data. In contrast, JSON-LD is another kind of RDF file which does not require an inference mechanism, making it more suitable for faster and simpler queries. Thus when working with JSON-LD, queries can be performed by Python Dictionary methods or RDFLIB Python package similar to SPARQL. These approaches allow for data filtering through specific conditions, making it an efficient way to access relevant information.

4) Host and test: Host the FAIRified datasets and metadata files. Test their interoperability and machine-readability using query techniques.

3.2 Data framework design

Through the above literature review, we obtained a foundational understanding of relevant research and algorithms. Based on the literature review findings, we further

examine the data frameworks used in the core papers [3][24][25] and critically analyze them against key standards, including multi-source data integration, support for multi-task learning, and potential for data reuse. Based on the identified limitations, we propose an optimized framework to address these gaps. Our framework should be built around two core components:

1) Multi-source data integration & Multi-task ML support: We need to ensure comparability and compatibility with diverse acoustic data sources integration through a modular pipeline capable of processing datasets in varying data structures and formats. By standardizing the data, the framework enables acoustic calibration and seamless application of machine learning models.

2) Cross-team data reuse for collaborative tasks: To facilitate data exchange and reuse across research teams, metadata file serves as a semantic bridge. Key results are stored in standardized, ontology-aligned labels for consistent updates. Because they are process-generated results, before going into the next step of the data life cycle, they need to be re-assessed and re-annotated. This validated metadata not only supports ML training, but also allows rapid extraction of anomalous events from lengthy soundtracks in downstream tasks.

3.3 FAIR data pipeline engineering and testing

In this phase we conduct the optimized framework through IT engineering, following the FAIR data workflow guideline. The main steps are as follows:

1) According to the data framework, design the data pipeline and define the technical stacks for engineering.

2) Use Python and machine learning algorithms in Pycharm IDE for building the pipeline. Load and split the long raw soundtracks into 60seconds flac files, and link the data with metadata as inputs. Each step is modularization as an independent python script so that it is easier debugging, testing, and future updates.

3) During the computational process, the metadata file is continuously updated - especially the machine learning labels, to record analysis results and support further data querying. Before data sharing with the downstream teams or the

analysis results going to the next round of the data life cycle, there should be an extra step for reassessment.

4) Once the Python pipeline is workable, translate the pipeline into Snakemake, a widely used workflow management system in scientific research. Define each pipeline step as a rule, with clear inputs, outputs, and processing code modules. It ensures the reproducibility across systems by setting up environments. Figure 6 shows the structure of a FAIR computational workflow. A workflow clearly describes how data moves and how different components like datasets, scripts, or machine learning models are executed and connected. These components should be modularized and can often be reused or recombined in other workflows, thereby promoting efficiency, reproducibility, and methodological transparency.

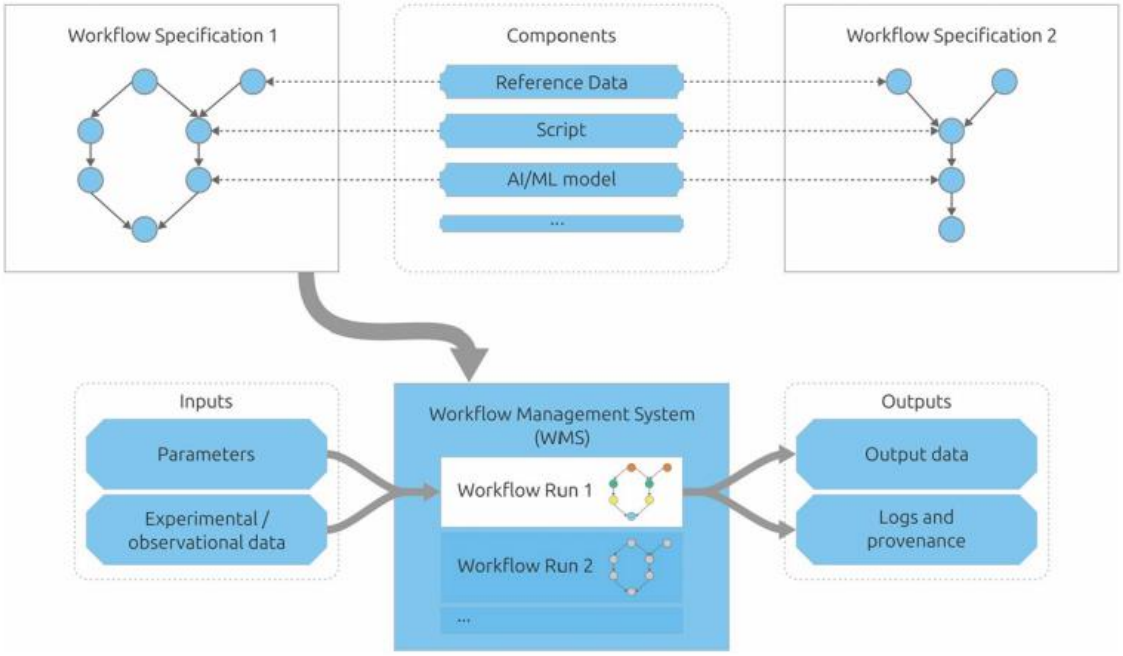

Figure 6: Example of a FAIR computational workflow [12]

5) After finish testing, upload and open the raw soundtrack data in Zenodo and the whole experiment documents in GitHub.

*A note on the main engineering tools:

Python: In this experiment, we chose Python, a widely used language for data analysis. It supports a rich ecosystem of libraries for big data processing and machine learning, and is compatible with formats such as JSON-LD, CSV, and various audio file types.

Snakemake: Executing a workflow involves standardizing this abstract specification using a workflow management system (WMS), which defines the execution by supplying necessary inputs (e.g., data files, parameters) and managing dependencies between tasks. As the workflow runs, each component is executed in order, producing outputs along with logs and metadata that record how the results were generated or updated [12].

Among available WMS tools, Snakemake is widely used in data-intensive domains such as bio-informatics and environmental data science. Snakemake enables the creation of scalable, automated, and reproducible workflows, described using a concise and human-readable syntax based on Python. Workflows are encoded in a 'Snakefile', where each rule specifies input and output files, the command or script to be executed, and optional resources or constraints. By handling job scheduling, dependency resolution, and parallel execution, Snakemake significantly reduces manual intervention and facilitates the development of robust and reuseful data pipelines. Comprehensive documentation and community support further contribute to its adoption in research and open science.

3.4 FAIR assessment

The FAIR assessment guideline provides specific sub-metrics to evaluate the FAIRification result. As shown in the Figure 2, each indicator is accompanied by a clear description, allowing it to serve both as an assessment tool and as a practical guideline for FAIRification. For example, F1 assesses whether data are assigned globally unique and persistent identifiers, while R1 focuses on whether data and metadata have rich description and accurate relevant attributes to be reused. The maturity model enables organizations and researchers to systematically identify gaps in their current data practices and prioritize improvements.

An automation tool for the ontology is applied called 'F-UJI', it is a web service to programatically assess FAIRness of research data objects at the dataset level based on the FAIRsFAIR Data Object Assessment Metrics. These Metrics are similar to the thought of FIP assessment tool suggested by Prof.John Graybeal. For the computational pipeline we use the FAIR maturity degrees to manually assess.

4 Related Literature

This chapter provides an overview of literature related to marine acoustic data, with a focus on key processing pipelines and machine learning models. Particular attention is given to data frameworks at the soundscape level, summarizing various preprocessing methods and ML algorithms relevant to our later-on design. Based on the review of 21 relevant papers, we synthesized common modeling strategies and algorithms across marine soundscape data machine learning workflows. (The full literature review process is shown in appendix A).

4.1 Why applying machine learning for acoustic data

The marine soundscape combines sound sources categorized as geophony, biophony, and anthrophony [29]. It is characterized by spectral, temporal, and spatial features that vary depending on the location and time [30]. In marine environments, the main biological sound producers include marine mammals, soniferous fish, and invertebrates. Soniferous fishes, a diverse group of vocal vertebrates, use acoustic signals for a range of social interactions and exhibit considerable variability in their life histories [14].

According to the systematic review by D.A.Nieto et al. [13], three main methods for analyzing ecoacoustic, bioacoustic, and soundscape data were identified: manual examination of acoustic events through listening to recordings or visually inspecting spectrograms; the use of acoustic indices to summarize variations in acoustic energy; and automatic recognition of sonotypes using machine learning algorithms. In their study, acoustic indices were commonly used, including the Acoustic Complexity Index (ACL), Acoustic Entropy Index (H), Normalized Difference Soundscape Index (NDSI), Acoustic Diversity Index (ADI), Acoustic Evenness Index (AEI), Bioacoustic Index (BI), Spectral Entropy (Hf) Index, and Temporal Entropy (Ht) Index etc. These indices are frequently applied in biodiversity assessments and provide valuable insights into the acoustic characteristics of ecosystems, helping researchers understand environmental changes, biodiversity patterns, and species behaviors. However, the use of acoustic indices comes with certain challenges. They can be sensitive to noise, require expert knowledge to select the most suitable indices for a particular study, and there is no

consensus among researchers on how to interpret them. Additionally, these indices may vary in different ecosystems [13].

Due to the limitations of using sound indices alone, some studies have started to incorporate ML as an additional analytical tool. Giuseppa Buscaino's research also indicates that analyzing long acoustic data series is challenging because each ecosystem has a unique soundscape signature that changes over time, such as variations within different times of the day, across seasons, or within multidimensional spaces. For better understanding to the ecosystem, collecting more data and improving automatic analysis, as well as other applying data analysis techniques are necessary [15].

ML has the potential to revolutionize PAM for ecological assessments [16]. Since 2008, ML methods have increased rapidly [13]. A more mature application of ML can be seen in terrestrial fields such as bird studies [1], from which the marine field has also adopted models and data management methods. In the study by Williams et al. [17], they created a bridge between sound indices and machine learning. The combination of sound indices was used as input features, and a supervised learning approach was employed to train a regularized discriminant analysis (RDA) algorithm, which classified sound recordings into either healthy or degraded habitat categories. This study provides the first evidence that using compound indices along with machine learning can outperform the use of single ecoacoustic indices in a tropical reef domain. This approach has the potential for other usages in marine and terrestrial habitat applications.

4.2 Main ML Algorithms applied in marine acoustic data

For soundscape level, most studies in this field employ supervised learning algorithms such as CNNs and RNNs. Roca and Opzeeland [18] applied twenty-three distinct acoustic metrics to develop a supervised approach for distinguishing between two different acoustic environments, using Random Forest for feature analysis and classification.

However, unsupervised learning algorithms, such as K-Means, U-MAP, PCA and GMM, have recently became as new directions for acoustic data research. Vasudev P. Mahale et al. used K-Means and PCA to classify fish vocalizations [14]. Unsupervised

approaches can also be used to visualize and explore ecosystem data in feature space, directly predict ecosystem health and monitor it longitudinally over long periods in autonomous monitoring systems [24]. In the study by Sarab S. Sethi et al. [24], a combination of CNN for embedding and U-MAP was used to directly assess ecosystem health, and a GMM model was employed to automatically monitor anomalous sounds, enabling real-time monitoring.

In marine aspect, unsupervised learning was mainly use in classify if the coral reefs and fishes have high or low diversity level by sound records, and to classify the overall landscape biodiversity level [3][25]. There are several promising future research directions in underwater acoustics using ML, including Physics-Informed Neural Networks (PINNs) for sparse data, transfer learning and domain adaptation for enhanced model generalization, ensemble and hybrid approaches to improve performance, active learning and data augmentation to address limited labeled data, and the development of explainable ML models for better interpretability and trustworthiness [26].

The downstream tasks in individual levels, machine learning applications in marine PAM data primarily focus on species identification, such as for specific whale [19], fish [20], and noise source detection like vessel identification [21]. However, these types of identifications require high-quality datasets and corresponding annotations, which demand significant time and labor. While methods like few-shot training [22] and transfer learning with pre-trained models can help alleviate some data scarcity issues, such as using bird models for coral reefs [16], the available marine acoustic data is still insufficient to fully meet the training requirements [23]. Although semi-supervised and unsupervised learning methods will become promising directions, with a growing trend toward label-free approaches that can handle the large volumes of data collected [13], mature solutions for individual tasks are more using supervised learning algorithms.

4.3 Review Findings

The use of unsupervised algorithms for soundscape level ecological assessment becomes a new trend, including K-Means, U-MAP, PCA and GMM, while more individual level tasks still rely on supervised learning models such as CNNs, Random

Forests, and RDA for species identification and event classification. Therefore, the design of our framework needs to accommodate the requirements of both types of tasks.

Among the reviewed literature, only three studies [3][24][25] fully align with our research focus, applying machine learning methods at the soundscape level, so we will mainly referring on the three studies when designing our pipeline.

Additionally, these machine learning tasks rely heavily on the availability of sufficient high quality data. However, most of the reviewed paper focus on algorithm development, with relatively limited attention paid to the data management, interoperability, or reusability of the datasets used. Therefore, effective data acquisition and sharing, along with the establishment of standardized data structures across organizations are crucial. By using ontologies to unify data features and contextual indicators, research efficiency can be significantly enhanced.

5 Marine Data Source and Ontology Result

To answer RQ 1, 'How can marine soundscape data be integrated, through ontology design to better support machine learning requirements?', we combine the findings from the data and ontology review with the results from the interviews. Based on this, we carry out the data FAIRification process.

5.1 Pre-FAIRification results: Data source and ontology review

5.1.1 Public data sources and ontologies

To inform the ontology design and assess existing integration gaps, we first examine available public data sources and ontologies relevant to marine soundscape research. The open databases are mainly in two categories: 1) broad biodiversity databases that include marine or ecoacoustic data, and 2) specialized databases focusing on marine or exactly marine acoustic (PAM) data. GBIF and OBIS are two popular universal biodiversity databases:

GBIF: Global Biodiversity Information Facility (GBIF) is a United Nations-sponsored global platform for sharing biodiversity information, providing free and open access to species and ecosystem data from around the world. It is based on the DwC standard and supports domain-specific extensions such as classification, occurrence and environmental measurements. GBIF emphasizes interoperability and ontologies such as biodiversity and environmental metadata. Updated to January 2025, related extensions registered here such as 'Audiovisual Media Description', 'Simple Multimedia', 'EOL Media Extension 1.0' just simply mentioned the data type might includes sound, and the 'Extended Measurement Or Facts' add more metadata on the measurement and environment background information, without focus on marine acoustic data design (*https://rs.gbif.org/extensions.html)*.

OBIS: With similar data types across marine species and ecosystem data, Ocean Biodiversity Information System (OBIS) works closely with GBIF, and it has the innovation that using the Event and Occurrence Core, with 'MeasurementorFact' and 'extendedMeasurementOrFact' extensions to make sure more information is included (*https://manual.obis.org/formatting.html#extensions-in-obis*).

In addition, several regions and countries, including Europe, USA, Australia, maintain their open ocean data systems. Among these, the following platforms include underwater PAM data:

IOOS-ONMS (USA): A key program under National Oceanic and Atmospheric Administration (NOAA). As part of NOAA's data ecosystem, it integrates observational data into NOAA portals like NOAA's Big Data Program and NCEI (National Centers for Environmental Information). IOOS also connects to global systems such as GOOS (Global Ocean Observing System) and has strong interoperability with OBIS. IOOS metadata standards are primarily based on the Darwin Core (DwC) vocabulary for biodiversity data and the Climate and Forecast (CF) metadata conventions and Attribute Convention for Dataset Discovery (ACDD) for netCDF formats. These are often supplemented with IOOS-specific extensions to meet ocean observation requirements.

Beyond general marine data, IOOS provides specialized marine PAM (Passive Acoustic Monitoring) datasets and clear definition for PAM metadata with project mission & platform & recording equipment information. It is one of the leading data platforms for PAM data, and widely used by researchers. Under IOOS category, the exact PAM data project is called NOAA's Office of National Marine Sanctuaries (ONMS).

EDMED and AtlantOS project (Europe): Both focus on Europe marine, EDMED focus on European marine data systems, designing ontologies like EDMED for interoperability. In Europe, EDMED (European Directory of Marine Environmental Data) serves as a metadata catalogue developed under the SeaDataNet infrastructure, aiming to improve the interoperability of marine environmental datasets across European research institutions. EDMED uses standardized metadata schema (e.g., ISO 19115, SeaDataNet vocabularies) to describe datasets. While AtlantOS is more on Atlantic Ocean hydrographic information, without any acoustic data publication. It is a broader initiative aimed at building an integrated Atlantic Ocean Observing System, aligning regional efforts with the Global Ocean Observing System (GOOS).

Soundlib (Netherlands): The Soundlib project, led by VLIZ, aims to create a FAIR underwater sound library of the North Sea for machine learning applications. It will collect and annotate long-term recordings, develop scalable database architecture, and use ML to efficiently classify and analyze sound events, supporting marine ecosystem

monitoring, ecological research, and noise impact assessment. But till the date of April 25th, 2025, this dataset is not published yet.

PANGAEA (German): An open-access repository for general environmental data, supporting various biodiversity-related ontologies. It includes the underwater PAM data we need. Data could be select by project, location, depth etc, and within the dataset detail page, platform provides a clear visual spectrograms for discover the sound data.

NERC (UK): National Environmental Research Council (NERC) manages marine and other environmental data in the UK, and it provides access to ecosystem-focused controlled vocabularies such as NERC Vocabulary Server (NVS).

AODN (Australian): The Australian Ocean Data Network provides open marine data access, focusing on regional marine ecosystems, such as the Great Barrier Reef and Southern Ocean. The platform is well organized, and with rich filters including project, location, depth range and very detail metadata file. Many datasets are accompanied by detailed metadata conforming to international standards (e.g., ISO 19115).

On the ontology aspect, there are some mature ontologies on marine ecosystem and sound measurement aspects, the major marine public databases are organized by the standardized ontologies, with unified vocabulary in metadata. Darwin-core is a foundation standard for sharing biodiversity data. Under this standard, there are some data sources and ontologies on marine (acoustic) data as table4. Ontologies focus on different aspects. For example, IOOS has PAM-specific vocabulary, but others like MMISW, NERC are more on general marine terms, OBIS's is more on biodiversity terms, and SSN is on general sensors' terms. And for datasets, some are aligned with their own organizations' ontology extensions, such as IOOS and OBIS, but some are limited.

Darwin Core (DwC) shares conceptual and practical similarities with the Dublin Core Metadata (DC) standard. As a result, DwC is built upon the Dublin Core Metadata standard, and should be seen as DC's extension for biodiversity applications. Maintained by the Taxonomic Databases Working Group (TDWG), its latest version was released on 2023-09-18. According to its scope, DwC focuses on four aspects: collections of biological objects or data, terminology for biological collection data,

compatibility with other biodiversity-related standards, and facilitating the addition of components and attributes for biological data

(*https://www.tdwg.org/standards/dwc/#scope-of-darwin-core*). Among 1,166 terms, only 3 terms relate to acoustics (keywords: 'sound' and 'audio'), and none directly describe sound. Therefore, creating extensions for emerging PAM technologies is essential to address applications and collaborations.

Some global open ocean data sources, including GBIF and OBIS, have made their own extensions on biodiversity based on DwC and environmental knowledge, and some platforms provide extension registration service for public. The summary of data sources and their ontologies is as following Table 4:

Data Source	**Region**	**Ontology** **Standard**	**Focus**	**Notes**
GBIF	global	DwC	General	
OBIS	global	DwC	Marine	
IOOS-NOAA	US	IOOS Metadata Profile 1.2 (ISO 19115, NOAA NCEI NetCDF, ACDD1.3, CF1.7)	Marine	Rich marine data, the ONMS project is focusing on PAM
EDMED	EU	ISO 19115, SeaDataNet vocabularies	Broad Marine data	
AtlantOS	EU	various	Atlantic Ocean hydrographic information	No acoustic data
Souhlib	NL	Unknown	Marine Sound	A project, but the data cannot be found yet
PANGAEA	GR	Various (DwC, ISO 19115, CF and other extensions)	Marine	Have some PAM sound, selected in the experiment in our study
NERC	UK	NVS, various	Marine	
AODN	AUS	ISO-19115-2	Marine	Have some PAM sound, selected in the experiment in our study

Table 4: Summary of marine and acoustic data source and their ontologies

5.1.2 Relevant ontologies in our research

We also review widely-used environmental and biodiversity ontology portals for term searching. For example, EcoPortal is a global repository integrating various ontologies for ecological data, while BiodivPortal adopts the Darwin Core (DwC) standard, widely used for biodiversity datasets. EcoPortal is more general on ecological aspect, and BiodivPortal is frequently used by above data platforms and has more connection with DwC. These portals provide valuable references for aligning our acoustic metadata model with existing standards and practices. In this research, we will use BiodivPortal as the searching platform for vocabularies. Because our research requires 4 aspects of metadata, including metadata in general basic description, Marine, PAM acoustic and Machine learning, so here the combination of multiple ontologies terms is required as Figure 7.

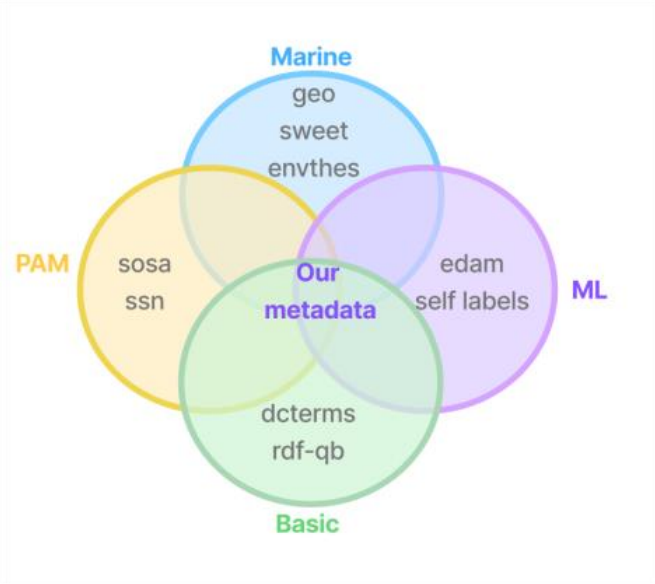

Figure 7: The sources of reference ontologies

Main relevant ontologies	Focus	Usages
DwC	Core Biodiversity terms	The foundational terms of biodiversity
RDF-QB	Statistical structure for RDF	Some basic attributes for RDF files
ISO 18405:2017	Geography	The global standard to describe geographical terms (used in US and AU, but here we adopt GEO instead)
GEO	Geography	Similar as above, but XX different
SWEET	Environmental semantics	Some semantic terms to describe environmental, Earth system, and observational processes
ENVTHES	Environmental thesaurus	Some environmental terms for supplement
SOSA	Sensor observation (core)	Some acoustic domain terms
SSN	Sensor system (extended)	Some acoustic domain terms
EDAM	Big data management	Focuses on big data topic; useful for describing workflow elements, contains machine learning
MAD	ML labels and other self defined entities	To have entities not in the above ontologies for our study

Table 5: Summary of different potential relevant ontologies

To support the integration and reuse of marine soundscape data for machine learning applications, a range of ontologies have been adopted or extended. Basic biodiversity and geospatial information is typically represented using Darwin Core (DwC), ISO 18405:2017 and GEO vocabularies. We adopt these mainly for ecosystem and spatial descriptions, forming the foundation for ecosystem-related datasets.

For sound recording aspect, underwater acoustics ISO 18405:2017, and the Semantic Sensor Network (SSN) ontology are for describing sensors and their observations, the involved procedures, the studied features of interest, the samples used to do so, and the observed properties, as well as actuators, based on DOLCE Ultra Lite (DUL) skeleton.

Additionally, since ML is in a different domain from environmental and biological sciences, certain specific terms also required, such as EDAM and self-defined ontology. As model inputs, these extensions provide essential contexts, including related environmental features, event time, event duration, event classification/label and

causation, or even textual description. As model outputs, the evaluation numerical results or classification labels, are valuable for tasks such as supervised or half-supervised learning, experiment reproduction, and the use of union metrics for comparing results. WORMS [10] to label the marine species for sound events.

5.2 Pre-FAIRification results: Interview insights

In this study, we conducted two interviews: one focused on sound data sources and machine learning algorithms, and the other explored marine ontologies and FAIRification techniques. Here are the key insights extracted from the two interviews:

Insights from Interview 1 (see details in Section 5.2.1):

Data source such as NOAA and Tethys could be considered to implement in the further study. We will use PAM data in NOAA data platform in our further experiment. The individual-level machine learning models, typically based on supervised learning, rely heavily on accurate labels for both training and storing results. So when we design the metadata, relevant labels should be included. Since these individual-level models are often downstream tasks following broader soundscape-level analyses, researchers prefer to work directly with raw acoustic data. Therefore, in addition to numerical metadata, access to the original sound data should also be provided and clearly linked.

Insights from Interview 2 (see details in Section 5.2.2):

Widely used ontology related to our research including Darwin Core (DwC), EnvThes, SWEET, WoRMS, SSN, SOSA etc., are validated to be useful with high quality for our research. We will adopt most of these ontologies into our ontology design. Use ontology recommender and annotation: In BioPortal or BiodivPortal to search best match ontologies and terms, these tools help bridge the gap between domain-specific terms and available semantic resources. For example, new topic like 'Machine learning', could be searched in the recommender, to find the most relevant ontology.

To build and manage an ontology, start by listing key terms aligned with metadata standards, track mappings to existing ontologies, and ensure long-term accessibility (even if marked obsolete), by publishing it in public repositories like GitHub. We will follow all these steps in the guideline to build our ontology. Also for the metrics, tool

like FIP according to the FAIRness standard, could be used in our final workflow assessment.

5.2.1 The first interview on Dec 18th, 2024

This interview was conducted at the Naturalis Biodiversity Center in Leiden. Two interviewees, Dr.Burooj Ghani and Dr.Vincent Kather, are researchers focused on long-term evolutionary studies and species-level classification using machine learning models trained on acoustic data, including bird sounds and those of other species. The one-hour interview was documented through detailed notes, from which we extracted key insights regarding data sources, metadata usage, machine learning models, and real-world applications.

One commonly used source mentioned by Burooj was Xeno-canto (XC), a collaborative, open-access platform dedicated to sharing wildlife sounds globally. While XC primarily hosts terrestrial species data such as birds, it fosters a strong community where users contribute and discuss wildlife recordings.

For marine acoustic data, they primarily turned to NOAA platforms due to their structured organization and rich acoustic archives. They also mentioned Tethys, a database designed for organizing and storing underwater PAM datasets. Additionally, they referenced the Detection, Classification, Localisation and Density Estimation (DCLDE) workshops, which provide access to whale PAM datasets open to expert discussions. However, they noted that while some metadata were publicly available, access to the actual sound data often required direct contact with data managers.

Data integration and cleaning were described as highly time-consuming, with inconsistent formats and incomplete metadata being common issues. The interviewees emphasized that unified data and metadata formats would significantly streamline their work. Their primary focus in machine learning was species classification using models such as deep neural networks. As a result, they prioritized model accuracy, species taxonomy, and evolutionary insights over extensive metadata, since their supervised learning tasks typically required only the sound recording and its corresponding ground true label.

They commonly converted acoustic signals into spectrograms or histograms and then applied CNN-based models for feature extraction. However, when asked whether they reused existing spectrograms or feature vectors prepared by other researchers or

platforms, they expressed a strong preference for starting from the raw acoustic data. This was because processing methods and parameters often varied by study, and specific approaches were essential to meet the specific needs of their analyses.

5.2.2 The second interview on April 28th , 2025

This interview was conducted online by ZOOM, along with Joëlle Stocker, we interviewed Professor John Graybeal for one and a half hours, it was documented through video record, detailed notes, and shared articles. Professor John now works in GO FAIR US Office and San Diego Supercomputer Center. He led the project on the Marine Metadata Interoperability community in the early 2000s, continually focusing on semantic web and ontology in a wide range of domains. He also developed BioPortal and Cedar as the leading tools for FAIRification. From this, we extracted key insights on metadata and ontology in the marine and acoustic domains, the tools and platforms supporting FAIRification, and the development and evaluation of FAIRification workflow.

In our discussion, we started with how to identify suitable ontologies for our research. Professor John particularly highlighted several ontologies relevant to marine data, including NVS Vocabulary, Darwin Core (DwC), MMI, EnvThes, and SWEET. For marine species, WoRMS (World Register of Marine Species) is the main resource. For sensors and instruments, Semantic Sensor Network Ontology (SSN) and its lightweight counterpart SOSA are useful, though somewhat abstract. However, since no single ontology may fully satisfy all the requirements of a specific research project, it is essential to apply one's own criteria to assess what ontologies to investigate or adopt.

Ontology Recommendation and Annotation Tools: For researchers uncertain about which ontology to use but who have a list of terms or concepts to define, tools like BioPortal offer ontology search functionality. Originally focused on biomedical domains, BioPortal has expanded to support a wide range of fields. It provides keyword-based search linked to relevant ontologies, making it a powerful entry point. The broader OntoPortal ecosystem includes domain-specific portals such as AgroPortal (agriculture), EcoPortal (ecology), EarthPortal (earth sciences), and BiodivPortal (biodiversity) etc. When entering keywords related to our area of research, these portals use a recommender to generate a ranked list of relevant ontologies. Rankings are based on factors such as ontology coverage, acceptance, knowledge detail,

specialization. This recommender system is especially helpful for beginners to identify suitable ontologies quickly. At a more granular level, annotation tools allow users to search vocabulary terms and return the most semantically relevant concepts, facilitating better linkage and reuse of data and terminology across systems.

John also mentioned a project in development called 'OntoChoice', which aims to provide a more standardized and systematic guideline for selecting ontologies based on predefined criteria.

To build and manage an ontology: A practical starting point for developing an ontology is to compile a list of key terms or term categories relevant to the project. These terms are often part of the project's metadata, so to aligned with metadata standards for interoperability, such as Darwin Core, Schema.org, and NetCDF CF conventions is important. The SKOS template provides a simple approach for vocabulary definition and allows for easy conversion from an Excel sheet to an RDF schema. CEDAR is another valuable tool that supports structured metadata creation.

Throughout this process, it is important to track potential mappings to existing ontologies that include similar concepts. As the project becomes more rigorous, managing our own ontology usually becomes necessary to accommodate more specific and tailored terms. However, if maintaining the ontology long-term becomes unfeasible, it can still be valuable. You can mark it as obsolete, and it will still serve important purposes: As a reference for others seeking to understand or build upon your work; As a seed for future development by the community. The most critical aspect is ensuring the ontology remains findable and accessible, ideally by publishing it in a recognized ontology repository and on GitHub.

Metrics (rules) to evaluation the ontology FAIRness: To assess the FAIRness of an ontology, the OBO Foundry provides a thoughtful set of principles - especially relevant if using GitHub (see: OBO Foundry Principles in *https://obofoundry.org/principles/fp-000-summary.html*). Similarly, the FAIR Implementation Profile Evaluator offers a checklist that closely overlaps with OBO principles. It includes key questions such as:

- Is the vocabulary representation compliant with established semantic web standards (e.g., SKOS, OWL, RDF)?
- Does the vocabulary provide labels and definitions for all its concepts?

- Is the vocabulary openly maintained and accessible with rich metadata in a public repository?
- Does the vocabulary's public repository support persistent and resolvable identifiers, versioning information, and provenance tracking?
- Is the vocabulary endorsed or maintained by a community organization or standards body?
- Does the model integrate data from multiple sources seamlessly and define qualified relations between entities?

In this context, a FAIR Implementation Profile (FIP) mini-questionnaire is available to guide us through the process of creating our own FAIR Implementation and assessment.

5.3 FAIRification and Post-FAIRification results

Our FAIRification process began with identifying suitable data for the experiment. We then defined metadata and ontology, checked feature coverage, and linked individual data files with metadata. Finally, we completed the FAIRification by hosting and testing the resulting data pipeline.

Data Sources: Based on both internet searches and recommendations from expert interviews (summarized in Table 1), we selected three main data sources for this study:

1) The experimental dataset published by Williams et al. [25];
2) The NOAA-ONMS platform;
3) The PANGAEA data platform.

We reused a subset of the dataset published by Williams et al. [25] as the baseline data, since it had already been successfully applied to soundscape-level machine learning models in marine ecosystem research, directly aligning with our research goal. This dataset was collected in French Polynesia at a shallow depth of 10–15 meters in 2021 [25].

In addition, we added two more datasets for the purpose of data integration experiment. The NOAA-ONMS platform provides rich marine acoustic data, while the

German PANGAEA platform is one of European leading scientific data portal. These two additional datasets allow us to test the FAIRification workflow and machine learning models over a broader spatial range.

From the NOAA-ONMS portal, we selected PAM data recorded at Stetson Bank, a mid-shelf bank located 80 miles off the Texas coast. The recordings were taken in 2023 at a depth of 22 meters.

From the PANGAEA portal, we selected PAM data from the FRontiers in Arctic marine Monitoring (FRAM) project, which includes detailed deployment documentation. The recordings were collected in 2020 from a deep sea location at 805 meters.

The original sources for these three datasets are as follows:

Williams' [25]: *https://zenodo.org/records/10539938*

ONMS: *https://storage.googleapis.com/noaa-passive-bioacoustic/onms/audio/fgb01/onms_fgb01_20230714/metadata/ONMS_FGB01_20230714.json*

PANGAEA: *https://doi.pangaea.de/10.1594/PANGAEA.967512*

So our experiment dataset is in 3 clusters [onms, pangaea, williams]. These original PAM data sources differ significantly in file format and recording duration. ONMS data are 4-hour continuous FLAC files, PANGAEA data are 10-minute OPUS files, and Williams data are 1-minute WAV files. The sampling rates also vary across the three projects. To create our experimental dataset, we finally randomly selected continuous soundtracks of fixed durations from each cluster: 1 hour of ONMS data, 10 minutes of PANGAEA data, and 1 hour of Williams data for the experiment. Later on these soundtracks will be resampled to 1-minute clips, in a total of 129 sound clips in our experiment (60 clips of ONMS, 9 clips of PANGAEA, 60 clips of Williams' [25]). These 129 clips will be regarded as individuals to link our metadata into JSON-LD. *Additionally 2 5-second anomaly sound clips as 'aodn' cluster are also added when testing the anomaly detection model (1 is whale sound and 1 is vessel noise.)

When looking into the metadata files of these 3 data sources (Figure 8), there are huge differences in the whole ontology logic, e.g. structure, entities, units and complexity. Overall, the metadata differences are in 5 aspects:

- Different metadata structural and design logic: Each dataset follows a totally different metadata structure and logic, including how the metadata is organized, the hierarchy of entities, and the way fields are distributed.
- Different metadata file formats and level: ONMS provides metadata in JSON format, PANGAEA uses TOML, while Williams' only includes dataset-level metadata that describes the dataset's background, without detailed data content metadata.
- Semantic differences in terms: Even when referring to similar concepts, the datasets use different terms or labels or even not controlled vocabularies. For example, they may all describe 'location' but use different field names or structures to do so.
- Different coverage of metadata fields: ONMS and PANGAEA include relatively complete metadata, covering most fields about the PAM data, whereas Williams' [25] lacks detailed metadata, with many fields left empty and not mentioned in the thesis.
- Differences in value representation and units: The datasets also differ in how they represent values, including units, data types, time format etc.

To solve these problems, first of all our task is to identify and align the key metadata terms across the different metadata files, defining our own ontologies, and then conduct the mapping rule to unified metadata.

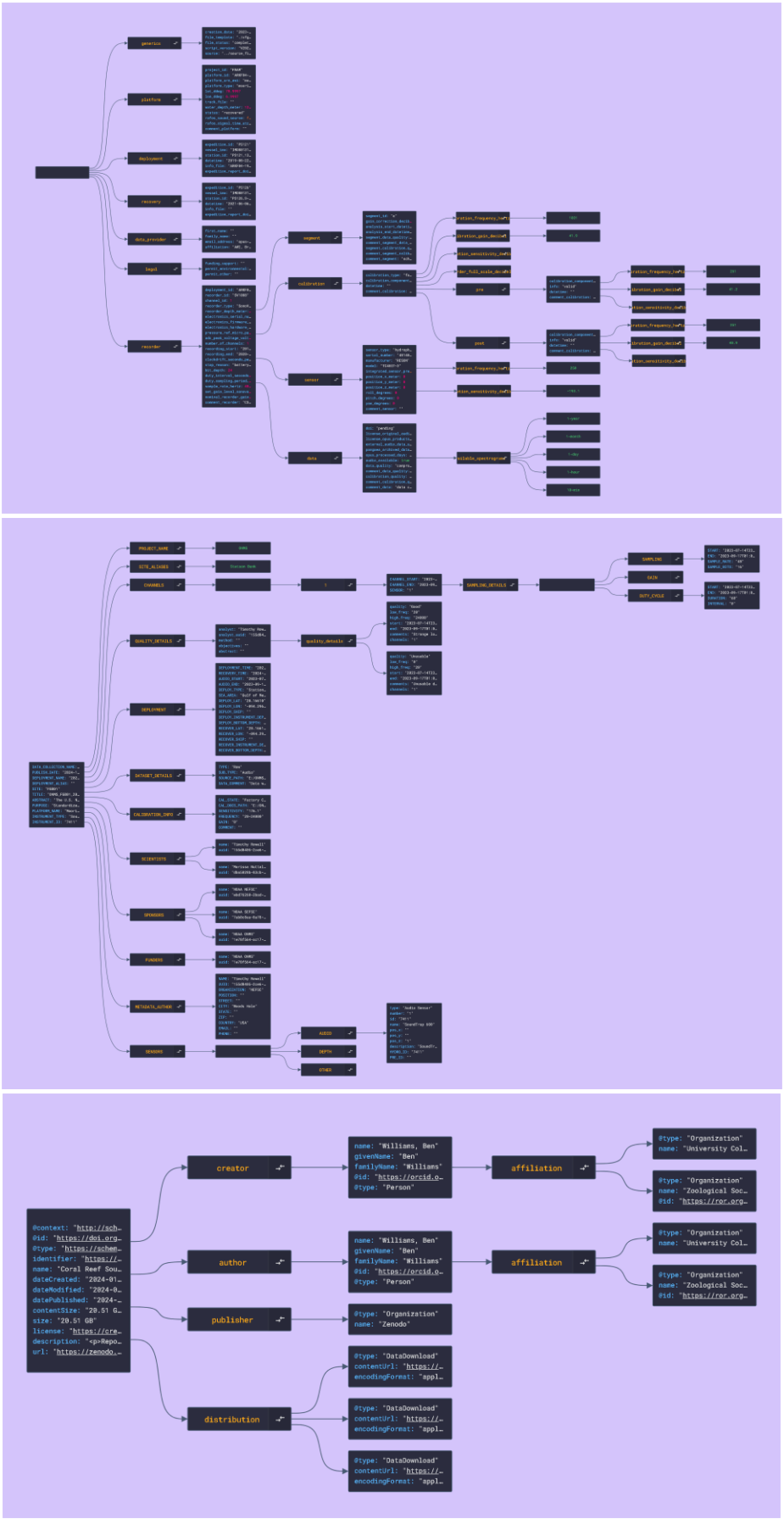

Figure 8: 3 datasets in 3 kinds of metadata formats

Define our metadata & ontology: Following the building ontology guideline from the second interview, we first list all the terms needed. Then categorize these terms into several domains, and search for suitable ontologies accordingly. From the existing marine PAM data ontologies suggested, Darwin Core (DwC), EnvThes, SWEET are used for marine ecosystem domain. For marine species, World Register of Marine Species (WoRMS) is the main resource for event description. For sensors and instruments, Semantic Sensor Network Ontology (SSN) and its lightweight counterpart SOSA are used.

For some new or uncertain terms, searching in BiodivPortal (one of the ontology portals also recommended in the second interview) is a method to find the suitable ontology. For example, when mentioning the ML concept, we search the keyword 'machine learning' in the searching box on the BiodivPortal homepage, and it suggests that the term 'Machine learning' in the EDAM ontology could be the best fit.

Combining these terms, we define our ontologies based on the above four aspects (Figure 12), and reusing the terms in these main existing ontologies. Shown in Table 6, here we classify the data properties for every soundtrack into six categories, including project, platform, deployment, ML, sample and soundtrack information. In each category there are some data properties for further hosting values. The terms are ensured the alignment with Semantic Web, using the prefix such as 'decterms', 'sweet' to clarify to source of the terms. We build this tree-shape ontology in Protégé, and export the TTL file as the whole definition of the ontology. Here, we use Protégé instead of tools like SKOS or CEDAR as suggested by expert, because Protégé offers greater flexibility for building ontology with hierarchical structures, supports visualization, and provides better control and integration with the Semantic Web.

To combine the three data sources with three metadata formats into our own ontology design, we need to also conduct a map rule table as table 6.

	Class	Data Property	ONMS	Williams	Pangaea	MAD Unit	MAD Description	Ref Ontology
1	project (sweet)	mad:project_name	project_name	name	project_id	/	the research project name	/
2	Platform (sosa)	dcterms:Location	SITE_ALIASES	/	/	/	the data collection location	DC
3		geo:lat	DEPLOY_LAT	/	lon_ddeg	degree	latitude	Basic Geo (WGS84)
4		geo:long	DEPLOY_LON	/	lat_ddeg	degree	longtitude	Basic Geo (WGS84)
5	Deployment (ssn)	sweet:Depth	DEPLOY_BOTTOM_DEPTH	/	recorder_depth_meter	meter	Deployment depth of the sensor	SWEET
6		sweet:StartTime	AUDIO_START	/	recording_start	/	record start time, in YYYYMMDD format	SWEET
7		sweet:EndTime	AUDIO_END	/	recording_end	/	record end time, in YYYYMMDD format	SWEET
8		sosa:Sensor	SENSORS-AUDIO-name	/	sensor_type	/	sensor name	SOSA
9		ssn:Frequency	FREQUENCY	/	sample_rate_hertz	hertz	recording frequency rate	SSN
10		ssn:Sensitivity	SENSITIVITY	/	/	decibel	recording sensitivity	SSN
11		envthes:calibration	CALIBRATION_INFO	/	calibration	/	calibration infomation and parameters, omit here	ENVTHES
12	Machine learning (edam)	dcterms:Event	/	/	/	/	1 or 0 or null, if event occurs label to 1, no then 0	DC
13		sweet:Description	/	/	/	/	descript abnormal event, use WORMS entities if marine speices	SWEET
14		mad:fish_diversity	/	/	/	/	1 or 0 or null, 1 is High and 0 is Low	/
15		mad:coral_reef_diversity	/	/	/	/	1 or 0 or null, 1 is High and 0 is Low	/
16	Sample (sosa)	mad:data_quality	quality	/	data_quality	/	1 or 0 or null, 1 is Good and 0 is Bad	/
17		qu:Duration	DURATION	/	/	second	all sound samples are cliped to 60s	QU
18		mad:Slice_URL	/	/	/	/	1 min sound slice data storage URL	/
19		edam:URL	/	url	expedition_report_doi	/	data original source	EDAM
20		qb:Slice	/	/	/	/	1 min sound slice name	QB
21	Soundtrack (mad)	mad:id	/	/	/	/	number to index the cliped sound samples, e.g. 1,2,....	/

Table 6: Metadata definition and mapping rule table

This mapping table helps connect the different metadata structures from the three original metadata files to the standard format we designed in our ontology. It shows how each field in the source metadata matches the properties in our ontology. This makes it easier to keep the meaning consistent across different formats and supports better data organization, searching, and integration based on Semantic Web principles.

Once the mapping rules are defined, we can use Python code to help us efficiently process and align the metadata across datasets. Code scripts can transfer the unified metadata files format into JSON.

Check feature coverage: As ML requires labels to store different models' result, here we also list the labels as extension. Also considering the spatial & temporal distribution, and other geological information needed for better data storage, management and retrieval, these features are also included. So it is available and complete for the further data linkage as triples.

Individual data linkage with metadata: This process involves two steps:

First, we prepare an Excel spreadsheet where each row represents one sound clip individual, and each column corresponds to a data property defined in our ontology. Based on the mapping rules provided in Table 6, a Python script is used to automatically fill in this metadata grid. The script takes original values from the source metadata files and maps them to the appropriate classes and properties in the ontology. During this step, data values in the Excel file are also cleaned, including standardizing units and time format, filling in the missing values mentioned in the thesis, and handling missing values (which are set as 'None').

Second, the cleaned metadata Excel sheet is then converted into JSON-LD format. The resulting JSON-LD file consists of two major parts: A context section at the top that declares the ontologies being used and referenced; The main body, which contains the linked data - the one entry per individual soundtrack, aligned with the ontology structure.

As a result, each sound clip entry in the JSON-LD file includes all relevant properties under their respective ontology classes. The result is a unified metadata file where every resampled sound clip is described using semantically structured and interoperable metadata. An example of such a sound clip in the JSON-LD format is shown below in Figure 9:

Figure 9: An example of N0.49 sound clip in JSON-LD

When the JSON-LD file is uploaded to an online JSON-LD visualization playground (*https://json-ld.org/playground/*), the entire metadata structure can be visually explored as Figure 10. This tool allows users to clearly see how each sound clip individual is linked to ontology classes and properties, making the relationships, hierarchy, and semantics much easier to understand and verify.

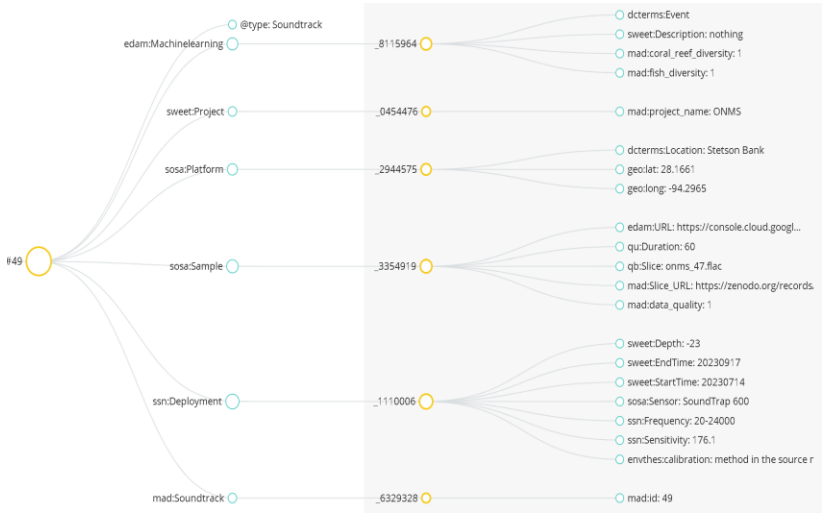

Figure 10: Example of Individual data linked with metadata JSON-LD file

*Trail on LLMs for metadata mapping: Additionally, we also experiment with a new approach to scale up metadata mapping using LLMs. For some metadata files, such as those from the popular PAM data portal AODN (Australian Ocean Data Network), the metadata is much longer and more detailed, often exceeding 1000 lines. To improve the efficiency of metadata integration, we test the use of large language models (LLMs), specifically GPT-4o-mini, combined with prompt engineering. The main idea is to provide the LLM with a prompt template that guides it to extract relevant metadata from the AODN metadata file and generate mapping rules aligned with our ontology. The output mapping example is as follows Table 7:

Metadata A Entity	Metadata B Entity	Mapping Description
@id	mdb:metadataLinkage/cit:CI_OnlineResource/cit:linkage	Maps the unique identifier of the metadata record to the online resource link.
@type	mri:MD_DataIdentification/cit:CI_Citation/cit:title	Maps the type of the soundtrack to the title of the data identification.
edam:Machinelearning	mri:abstract	Maps the machine learning context to the abstract description of the dataset.
sweet:Project	mri:citation/cit:title	Maps the project name to the title of the citation.
sosa:Platform	mdb:acquisitionInformation/mac:platform/mac:MI_Platform/mcc:MD_Identifier/mcc:code	Maps the platform information to the identifier of the acquisition platform.
sosa:Sample	mdb:contentInfo/mrc:MD_CoverageDescription/mrc:attributeGroup/mrc:MD_AttributeGroup/mrc:attribute	Maps sample information to the coverage description attributes.
ssn:Deployment	mdb:acquisitionInformation/mac:MI_AcquisitionInformation	Maps deployment information to acquisition information.
mad:Soundtrack	mdb:metadataIdentifier/mcc:MD_Identifier/mcc:code	Maps the soundtrack ID to the metadata identifier code.
qu:Duration	gex:EX_TemporalExtent/gex:extent/gml:TimePeriod/gml:begin	Maps the duration to the temporal extent of the dataset.
mad:Slice_URL	mdb:metadataLinkage/cit:CI_OnlineResource/cit:linkage	Maps the slice URL to the online resource link for accessing the dataset.
mad:data_quality	mdb:resourceConstraints/mco:MD_Constraints/mco:useLimitation	Maps data quality to the use limitation constraints of the dataset.

Table 7: Example of LLMs auto-mapping the metadata

The result is generally promising. While not all mappings are perfectly accurate, the LLM significantly reduces the manual workload of parsing and aligning complex metadata. It serves as a powerful assistant in handling semantic tasks at scale, with final verification and adjustments made by domain experts.

Host and test results: After generating the unified metadata file in JSON-LD format, we publish it via GitHub to ensure broad accessibility and long-term reusability. This open sharing approach allows researchers and developers to directly access, inspect, and integrate the metadata into their own projects, following FAIR data principles.

The 129 1-minute sound clips are compressed into a ZIP file for convenience and ~~is~~ uploaded to Zenodo, a trusted research data repository. Each sound clip is assigned to the same access link on Zenodo, which is referenced in the metadata under the property 'mad:Slice_URL'. This ensures that the metadata and data remain connected, while being modular and scalable for different use cases. (*Note: In our actual computational process, the ZIP file should first be unzipped so that the script can process each clip individually. Moreover, in a real federated scenario, the resource links (mad:Slice_URL) should be unique for each clip, allowing them to be queried and processed as separate federated individuals).

To verify the usability of the structured metadata, both data query tests including SPARQL querying and Python Dictionary querying work. For simpler usage, the JSON-LD file dictionary-based lookup and filtering is straightforward. This provides an efficient way to retrieve and process metadata in environments. Both methods prove functional and effective, confirming that the metadata can be flexibly queried and integrated into various research and analysis workflows.

After these steps, the data is FAIRified and ready for machine learning analysis.

6 Data Framework Design Result

To answer the RQ 2, in this chapter we point the key insufficiency in current marine soundscape level machine learning framework, and propose an optimized pipeline according to FAIR principles.

6.1 Current framework insufficiency

In this research, based on findings from the systematic literature review above, we mainly refer to three representative data processing pipelines [3][24][25] adopted in passive acoustic monitoring (PAM) studies of marine ecosystems. These pipelines are illustrated in Figure 11 and reflect a range of machine learning methods for ecological soundscape analysis.

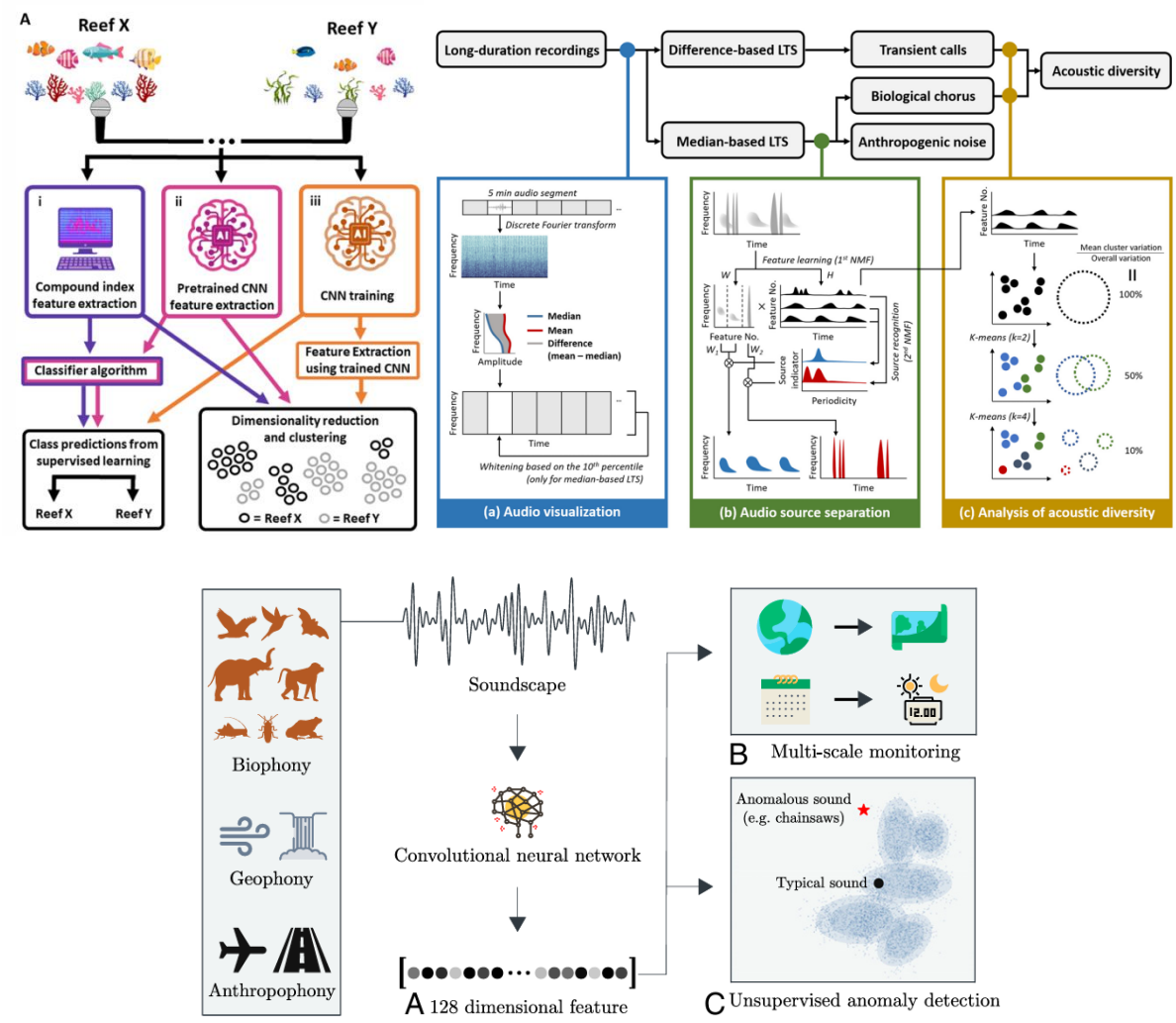

Figure 11: 3 main refer data pipelines (a. left, b.right, c.bottom)

Pipeline 1 (Figure 11a) [25] shows a best practice that begins with long-duration PAM recordings collected from multiple coral reef sites. The raw audio data is first transformed into a spectrogram using the Librosa library, a widely used Python toolkit for audio signal processing. This transformation is based on Fourier Transformation method, which converts the audio signal from the time domain to the frequency domain, allowing us to see how the signal's frequency content changes over time as spectrogram. For some data platforms like PANGAEA, it directly provides the pre-generated spectrogram for easy navigation of the acoustic data as Figure 12.

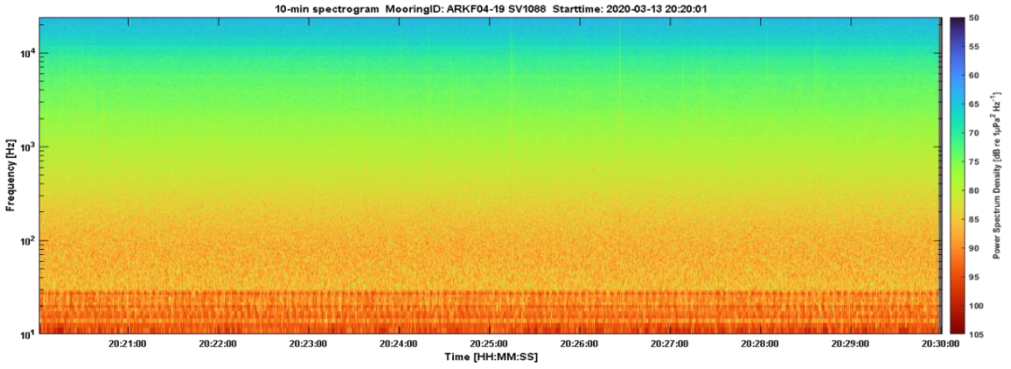

Figure 12: 10-minute spectrogram visualization in PANGAEA data portal

The spectrogram shows the distribution of acoustic energy over time (horizontal axis) and frequency (vertical axis). The color scale represents signal intensity, as warm colors indicating stronger intensity and cool colors indicating weaker, allowing a clear visual understanding of how sound varies across different frequencies over time.

The resulting time-frequency spectrogram is then fed into a pretrained Convolutional Neural Network (CNN), such as VGGish to extract high-level feature embeddings. Below Figure 13 explains the process of CNN, it uses a series of filters (convolution and pooling layers) to scan the spectrogram and extract key feature patterns into feature maps. These filters capture local features at different layers and gradually represent the spectrogram characteristics using numerical matrices. The final output is a compact feature vector that summarizes the essential attributes of the sound signal, making it suitable for further machine learning analysis.

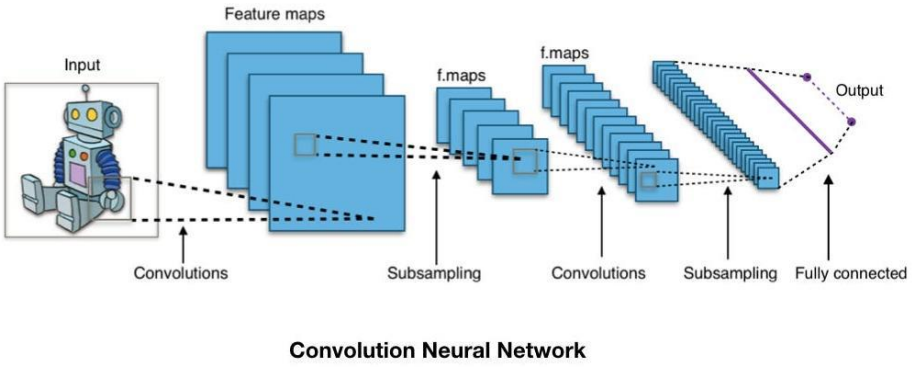

Figure 13: Convolution Neural Network Framework (CNN) [28]

These feature vectors are subsequently processed using dimensionality reduction technique (UMAP) and unsupervised clustering algorithm, enabling the classification and comparison of soundscapes for biodiversity level assessment.

Pipeline 2 (Figure 11b) [3] follows a partially different approach. It starts with spectral conversion through similar Fourier Transformations and optionally applies spectral whitening to enhance key frequency components and reduce bias from persistent background noise. From these transformed acoustic features, K-means clustering is used to cluster similar acoustic events and quantify diversity metrics. This pipeline is also designed for soundscape level data for biodiversity assessment as pipeline 1.

Pipeline 3 (Figure 11c) [24] shares methodological similarities with Pipeline 1 in its initial processing steps, including spectrogram generation and CNN-based feature embedding. However, it diverges in its end goal, utilizing Gaussian Mixture Model (GMM) instead of clustering for the purpose of unsupervised anomaly detection. This approach aims at identifying outliers or unusual acoustic patterns, which may correspond to rare biological events or anthropogenic disturbances (e.g., illegal fishing activities, sudden noises).

Together, these pipelines highlight the growing convergence of signal processing ML, and ecological analysis in modern ecoacoustics. Pipelines 1 and 2 are primarily designed for assessing biodiversity levels, while Pipeline 3 focuses on anomaly detection. This single-task-oriented design leads to two inefficiencies: There is a disconnect between soundscape-level modeling and downstream tasks such as species identification, behavioral analysis, and event detection. The downstream tasks can not reuse the results from previous training and separate the target data clips from sparse long data. Moreover, the lack of a modular and unified workflow framework makes it

difficult for researchers to reuse existing models and processes across different research objectives.Processes such as data preprocessing, feature extraction, model training, evaluation, and deployment are often implemented through scattered scripts, without standardized workflow encapsulation.

These two issues together shows that, although progress in task-specific modeling and data processing techniques, current machine learning research in soundscape analysis still lacks a unified, modular, and shareable workflow design. Open Science need to develop multi-task-oriented and FAIR machine learning workflows for ecoacoustic research, to improve flexibility in research design and the efficiency of data reuse. However, their application is often narrowly aiming at single-task objectives, and lacks a flexible, unified design capable of supporting diverse downstream tasks, like training species classification models.

6.2 An optimized data pipeline framework

To address the two key inefficiency identified above, insufficient standardized multi-task process design and insufficient FAIR computational workflow, the optimized pipeline (as Figure 14) is designed.

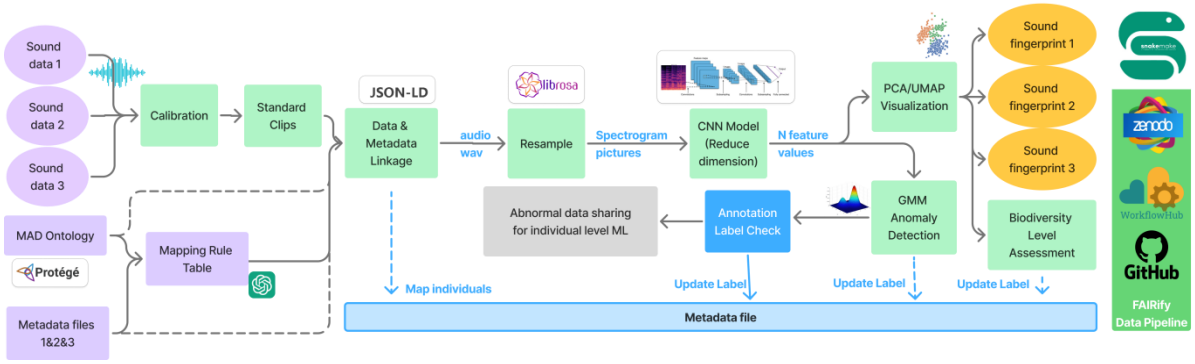

Figure 14: Overall data pipeline framework design

Common data prepocessing: Starting from calibration and clipping the long raw soundtracks into unified-duration clips, in parallel standardized ontology and metadata mapping are applied. The data and the metadata are linked into a semantic RDF format (JSON-LD) using our ontology schema. This step ensures interoperability and prepares the data for machine readability. LLMs could be used for efficient metadata mapping. Then, spectrograms generation via Librosa is followed by CNN-based feature extraction. This step standardizes acoustic data representation while

compressing information into feature vectors (N feature values) using dimension reduction technique, referenced in the research by Williams et al. [25].

*A special note: in this work, we did not implement a detailed calibration procedure. Rather, our approach designs as a prototype for a broader macro-level solution. We only applied a basic standardization of the resample rate using Librosa. But in real world applications, there are a variety of advanced calibration methods being actively researched and implemented.

Divergence in ML tasks: Once the standardized acoustic data vectors are obtained, they can be used in multiple ML tasks, including PCA/UMAP sound fingerprint visualization, biodiversity high/low clustering assessment using the K-Means algorithm, and GMM for anomaly detection. After anomaly detection identifies anomalies at the soundscape level, which can identify ecologically relevant events, these anomalies are explicitly labeled and updated in the metadata file, ensuring they are available for reuse in downstream ML tasks (e.g., training species classifiers on unusual events).

Metadata file as a semantic bridge: The key here is a metadata-driven method. A metadata file is created, links data clips, and is updated throughout the process, especially with the biodiversity level clustering results and the anomaly detection result labels. Therefore, for downstream individual classification tasks, sound clips can be reused directly by querying for the 'Event' label set to 1, as in <[dcterms:Event] equals '1'> in the metadata.

But as we mentioned before, the labels under the ‘edam:Machine learning’ class are all updated by the machine learning pipeline or observer's annotation, so they are not provenance features. Before the downstream teams query by these labels, these labels should be validated and could be regarded as changeable labels.

Finally, the pipeline is ready for computational workflow FAIRification with a series of tools. This pipeline ensures the outputs from one soundscape-level task, particularly anomaly detection, can be flexibly reused in downstream tasks like species recognition or event classification.

7 Data Pipeline Engineering and Testing Result

In order to answer the RQ 3, 'how does the optimized pipeline perform driven by the data models?', this chapter presents the construction of our optimized pipeline through code engineering, accompanied by testing of machine learning results and continuous creation and updating of metadata files. Once the entire pipeline is workable, we apply FAIR guidelines to FAIRify the computational workflow for reuse in Open Science.

7.1 Pipeline Modules

The overall pipeline is divided into eight modular script steps, all written in Python 3.12. They will be executed by order:

1) 0_toml2json: This script unifies the metadata files from TOML to JSON format for further processing.

2) 1_clip_raw_sound: This script first clips long acoustic recordings into multiple 60-second FLAC sound clips. It also creates a table to store metadata values, where each row represents one sound clip sample and the columns correspond to metadata terms, so the metadata values could be added into the table in next step.

3) 2_add_metadata: With unified sound clips and the metadata table ready, this step applies the mapping rules (as defined in Table 4) to automatically insert metadata values into the unified table. During this process, data cleaning is also performed, including adjustments to units and time formats, and filling empty values with 'Null'. The output is a unified metadata table enriched with cleaned metadata values.

4) 3_map_individuals: This script transforms the unified metadata table into a machine-readable JSON-LD format for interoperability. It has two main parts: first part uses the TTL ontology file generated in Chapter 5 with Protégé to extract hierarchical relationships and semantic web linkage information into the '@context' section of the JSON-LD. The second part, '@graph', represents each sound clip entry using a dictionary structure aligned with the ontology. From this step onward, all sound clips are linked in the JSON-LD metadata file.

5) 4_query_individuals: This is the data querying script, supporting both SPARQL queries and Python dictionary-based queries. After testing some sample queries, the entire dataset is ready to be used as a rich source for further data engineering.

6) 5_clip2features: This step is key to transforming the sound clips into numerical vectors. First, each 60-second sound clip is split into 12 segments of 5 seconds. Then, for each segment, the Librosa package is used to convert the acoustic data into a spectrogram. Due to varying sampling rates across different data sources, we follow the method in the work of Williams et al. [25] and parameters to standardize all audio files to a sampling rate of 32,000 Hz using Librosa. This rate is lower than the commonly used 44,100 Hz (CD quality), which helps reduce data size while retaining sufficient audio detail, thus balancing quality and performance. Next, we download a pretrained CNN model ('SurfPerch_v1.0')[25] for spectrogram feature extraction, which reduces the high-dimensional spectrograms into 1280-dimensional vectors. This deep learning model was trained on a large dataset of annotated environmental audio. As a result, each 60-second sound clip is converted into 12 segments, each represented by a 1280-dimensional feature vector. These vectors are essential for subsequent machine learning tasks. They are saved in a new vector table, indexed by the sound clip ID to ensure traceability back to the original audio data.

7) 6_umap & 6.1_diversity: The UMAP script uses the UMAP algorithm to reduce the 1280-dimensional vectors into 2D, and visualize how the acoustic data features are distributed in the plane. A similar dimensionality reduction algorithm PCA, is also applied for cross-validation. We input the vector table here, and the script outputs the UMAP and PCA result figures as sound fingerprints. From these figures, we can make an initial judgment about the effectiveness of feature extraction and how well acoustic data from different sites can be separated. In parallel, we can also train a clustering classification model for biodiversity assessment (6.1_diversity script). Since we only classify the level as low or high, we use the K-Means algorithm with the number of clusters set to 2. It outputs a clustering result figure, and assigns each sample an output of 0 or 1, representing different biodiversity clusters. But we don't know the biodiversity levels of the three datasets, and the result requires at least one known label to classify the others, this step is shown only as an example and is not implemented into the pipeline.

8) 7_gmm: Once the feature extraction works well, we can use the GMM algorithm for anomaly detection. We input the vector table and the metadata file. Using the GMM algorithm, it identifies anomalies based on differences in the Gaussian distribution of the vectors, selecting all log-likelihood values below the 2% percentile as the threshold for anomaly detection. For the detected anomalies, the corresponding clips in the metadata file will have their 'Event' label updated to 1 (indicating anomaly), while the normal clips remain unchanged.

7.2 Pipeline testing results:

To evaluate the effectiveness of the machine learning data pipeline, we conduct a series of tests using three sample datasets, and also 2 whale call 5-seconds sound data are added as known anomaly clips. The results of feature extraction, clustering, and anomaly detection functions are workable with the summary below. (*Please note that in this experiment we did not conduct a detailed calibration research, so the following results show more about the functional workability of the whole data pipeline rather than the actual modeling effects).

7.2.1 Feature Extraction

Figure 20 below demonstrates that the feature extraction process (as from script0 to script6) is successful, which is capable of clearly distinguishing the acoustic features of different sound clips. Specifically, both the UMAP and PCA dimensional reduction techniques are applied to the 1280-dimensional feature vectors, projecting them into two-dimensional space for visualization.

In the UMAP plot (Figure 15a), we observe well-defined clusters, indicating that the model has effectively captured patterns in the acoustic data. Similarly, the PCA plot (Figure 15b) confirms the effectiveness of the extracted features. Both plots shows that sound clips originating from different sources are well-separated in the reduced dimension feature space. This separation validates the success of the feature extraction process and provides a solid foundation for downstream analyses such as classification and clustering.

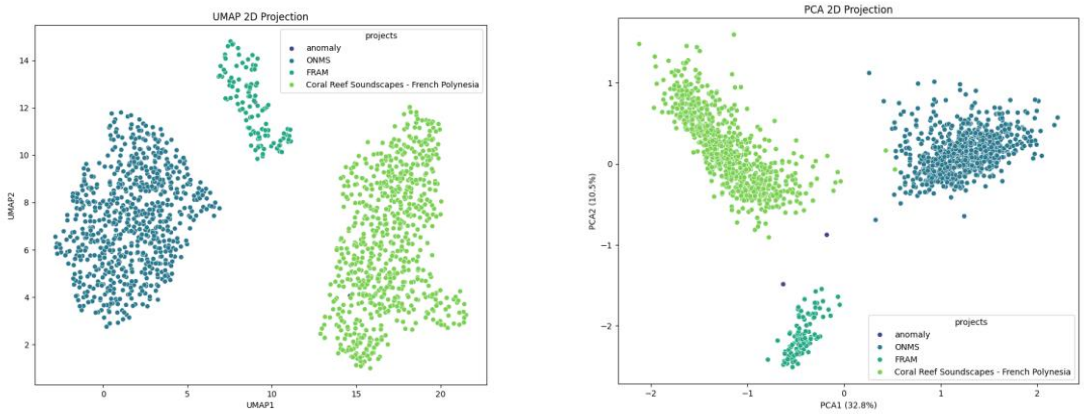

Figure 15: UMAP result(a.left), PCA result(b.right)

7.2.2 Biodiversity

To assess biodiversity levels based on the extracted features, a K-Means clustering algorithm was applied (as script 6.1). The goal is to classify the biodiversity level into two general categories (low and high), so the number of clusters is set to 2. The result is visualized in Figure 16, where each data point represents a 5-second sound clip segment, and the cluster labels (0 or 1) indicate their assigned biodiversity group.

As shown in the plot, the model has effectively divided the data into two distinct clusters (0 or 1), different colors represent different data sources. Although no ground-truth biodiversity labels are available for these datasets, this clustering result provides an exploratory perspective on potential biodiversity differences across soundscapes. With at least one known labels, this method could be extended for supervised classification or biodiversity index estimation.

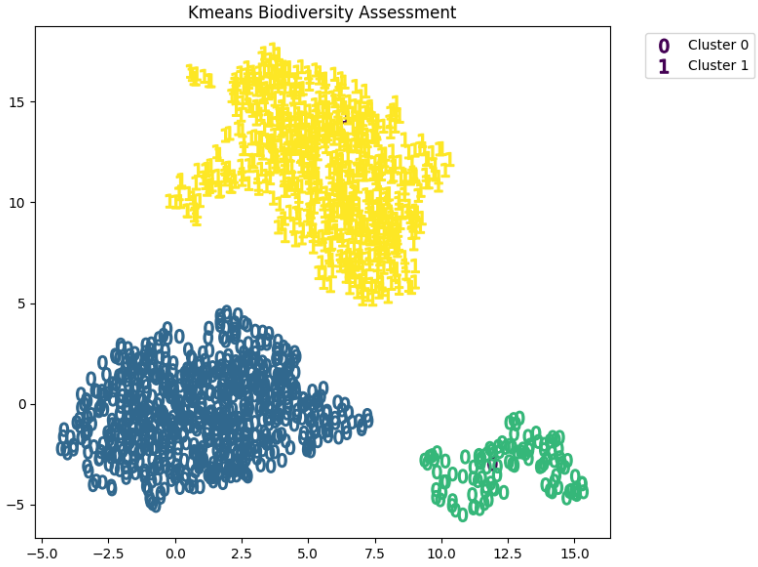

Figure 16: Biodiversity level K-means clustering

7.2.3 Anomaly detection

The Figure 17 below presents the successful results of anomaly detection using the Gaussian Mixture Model (GMM) algorithm (as script 7). In this two-dimensional visualization, each point represents an acoustic feature vector after dimensionality reduction. The purple dots indicate 'normal' sound clips, while the light blue dots represent abnormal as potential anomalies or 'events'.

From the plot, we could see that the normal data forms three distinct and dense clusters, which align with the expected Gaussian components learned by the model. In contrast, the abnormal dots (in blue color) are generally scattered and tend to appear at the outskirt of these clusters regions. This spatial distribution suggests that the anomalies deviate from the Gaussian distributions that represent the typical acoustic patterns. The GMM algorithm models the overall data distribution as a mixture of multiple Gaussian distributions. It calculates the log-likelihood of each point belonging to the learned distribution. This probabilistic approach allows for flexible detection of unusual patterns without requiring labeled anomaly data in advance.

Once these anomalies are identified, the corresponding clips in the metadata file are updated by setting the 'Event' label to 1 for anomalous, while retaining 0 for normal clips. This labeling is particularly useful for downstream applications such as ecological event detection and acoustic data quality control.

Listen back to the clips labeled as anomalous, we often find distinct sound characteristics such as 'clicks' or 'Da' sound, or sudden increases in sound intensity. These acoustic features might indicative of biological or anthropogenic events that deviate from the ambient background soundscape. To further validate the effectiveness of the GMM-based anomaly detection, we deliberately insert two known anomalous whale call clips into the dataset. The two clips are successfully identified by the model as outliers, confirming its capability to detect different acoustic events. This validation suggests that the GMM approach is so far robust and flexible for different event features.

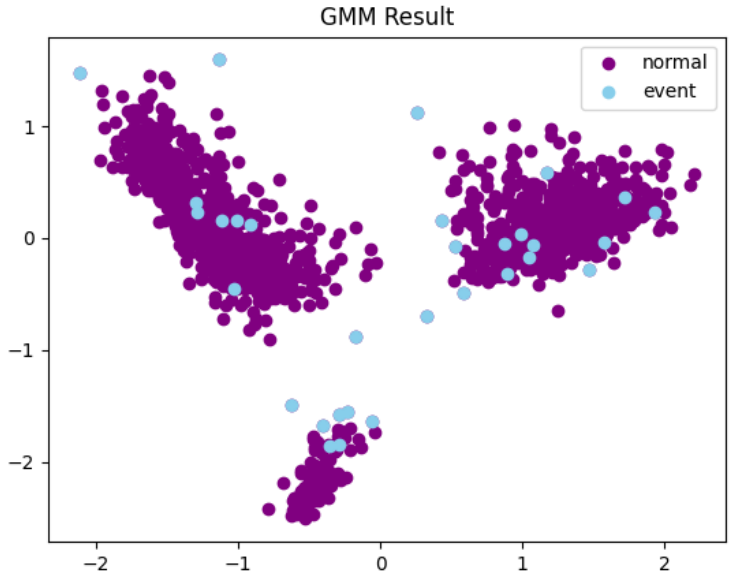

Figure 17: GMM Anomaly Detection result

7.3 Computational workflow FAIRification

As suggested by de Visser et al. [27], we have adopted the 10 tips provided as Figure 18 to implement FAIRification. Our solution uses various tools and techniques to ensure that the computational workflow aligns with the FAIR principles as below:

Findability:

- Workflow Registration: Register the workflow on platform WorkflowHub to make it findable and citable.
- Rich Metadata Description: Enhance the findability of the workflow by adding structured metadata using RO-Crate (Research Object Crate). Also when

uploading to Zenodo, the platform also asks for various description and generates the metadata automatically.

Accessibility

- Open Source Code: Store the source code in public repositories GitHub to ensure its accessibility.
- Provide Sample Data: Use the Snakemake workflow management system to include sample input data and results for better understanding and validation.

Interoperability:

- Standardized File Formats: Adopt standard file formats like FLAC, CSV, and JSON-LD to ensure smooth data exchange between different systems.
- Workflow Portability: Use Snakemake to enable the portability of the workflow across different computational environments.

Reusability:

- Reproducible Computational Environment: Create a reproducible computational environment using Conda to ensure the workflow runs consistently across different setups.
- Default Configuration Files: Provide default configuration files in Snakemake to simplify workflow reuse.
- Modular Design: Implement a modular workflow with Snakemake to allow easy component reuse and customization.
- Comprehensive Documentation: Offer clear and concise documentation (Readme.md in GitHub and our thesis explanation) to ensure that others can understand and reuse the workflow.

[27]

Figure 18: computational workflow FAIRification technique solutions

Our solution effectively integrates various open-source tools and standardized practices to achieve FAIRification of research workflows, enhancing transparency, reproducibility, and collaboration in scientific research.

Snakemake, as the key tool, automates and manages the computational workflow execution, providing a clear visualization of the process through Directed Acyclic Graphs (DAGs) that illustrate task dependencies and execution order. Figure 19 shows the Directed Acyclic Graph (DAG) of a sound data workflow built using Snakemake. The figure on the left provides a simplified view of the core components of the workflow and their dependencies, while the right one shows a detailed display of the specific input and output file paths for each step.

The workflow is designed to process raw sound data, adding metadata, mapping individuals, extracting features, and ultimately generating UMAP dimensionality reduction visualizations and GMM clustering analysis results. Implementing this workflow requires 4 main steps:

1) Configure Snakemake locally.

2) Rewrite the input and output paths in the existing 8 scripts to match Snakemake's standard (e.g., 'sys.argv[1]').

3) Create a separate rule file to specify input/output and script execution, standardizing the process.

4) Test the workflow to ensure it passes and generates the DAG for clarity.

Once tested successfully, the workflow can be reused by simply opening it in the CMD environment and running the corresponding command. This process showcases the powerful capabilities of the Snakemake workflow management system: by clearly defining the input and output dependencies of each rule, Snakemake automatically constructs the execution DAG, ensuring the repeatability and traceability of the data processing. Each step is encapsulated as an independent rule, making it easy to maintain and reuse. The related files also uploaded in GitHub.

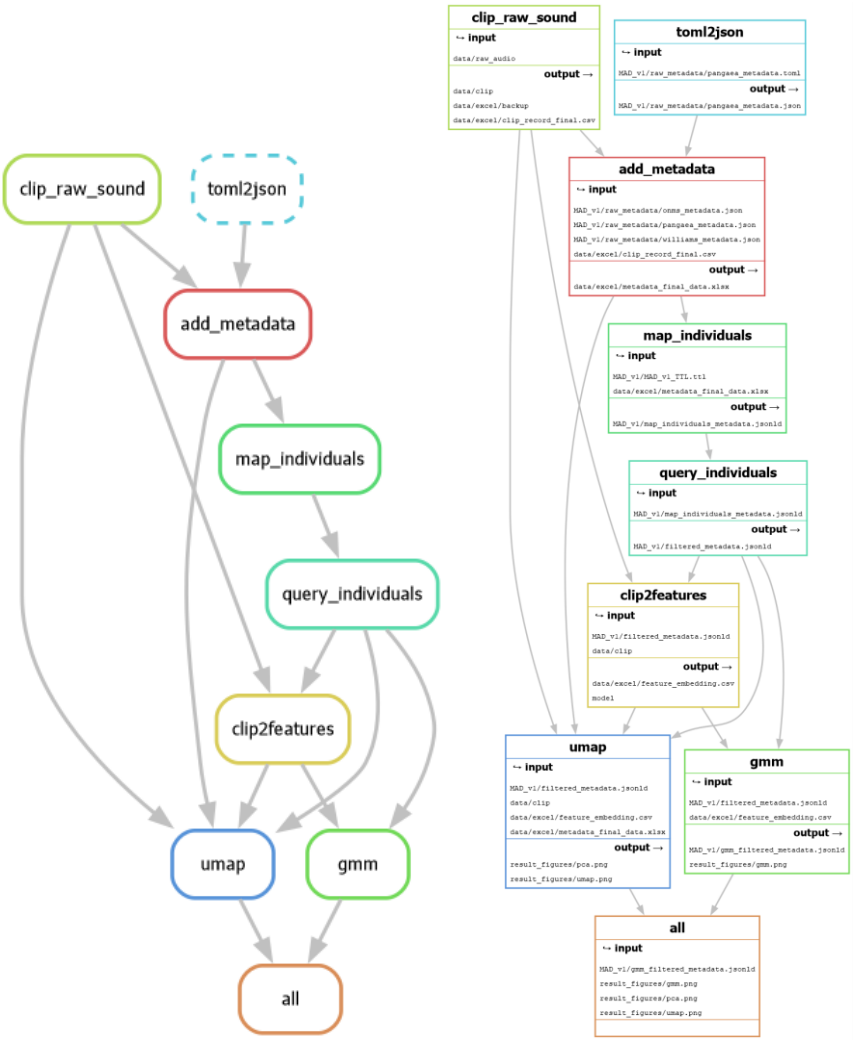

Figure 19: Workflow graph auto-generated by Snakemake

8 FAIR Assessment Result

For the final RQ 4, we conduct a FAIR assessment to evaluate the FAIRness in dataset and computational workflow level. The dataset shows a relatively high level of FAIRness, and the computational workflow is well aligned with the FAIR principles. Detail result is shown as 8.1 and 8.2.

8.1 Dataset FAIR assessment result

Here we use F-UJI, a web service to automatically assess FAIRness of research data objects (aka datasets) based on metrics developed by the FAIRsFAIR project. We select the latest metric v0.8 (based on the FAIR guideline in Figure 3), and input the Zenodo dataset URL link for assessment. Below Figure 20 is the summary of our dataset FAIRness assessment result:

Summary:

Figure 20: Our Dataset F-uji Score

Overall our dataset FAIRness is good, but according some criteria it could be improved in following points:

1) FsF-A1-01M-1 Information about access restrictions or rights can be identified in metadata: NO access information is available in metadata, thus it is unable to determine the access level.

2) FsF-I2-01M-2 Metadata uses terms from registered vocabularies that are identified by their name-spaces: No known vocabulary name-space URI is found which is listed in the LOD registry, mainly a list of URL links not registered in Semantic Web.

3) FsF-R1-01M-2 Information on the manner and form (file size and type or service (API) endpoint and protocol) in which data is delivered is provided: This part is missing.

4) FsF-R1-01M-3 Measured variables or observation types are specified in metadata: This information is missing.

5) FsF-R1.3-01M-1 Community specific metadata standard is detected using name-spaces or schemas found in provided metadata: This information is missing.

For points No. 2,4,5, we have already uploaded the JSON-LD file for further explanation, as the Zenodo metadata creation spreadsheet does not provide an option to fill in or upload this information. However, when the F-uji tool executes the assessment, it cannot automatically detect the JSON-LD supplementary metadata file in the dataset. There is a gap between the real FAIRness and the automated assessment.

For point No.1, Zenodo already provides an option for access control, and we have set it to 'open', but the tool ignores this information. For point No. 3, indeed, Zenodo does not offer this option either, so it could be further added in the JSON-LD metadata file as a supplement.

From the automatic assessment test, the tool's evaluation is generally efficient and accurate, but there are still some discrepancies in the details, the platforms and tools are not fully compatible with each other. Adjustments may be needed based on the specific circumstances of the project.

8.2 Computational workflow FAIR assessment result

FAIR metrics are more standardized at the dataset level, while the FAIR metrics for complete computational workflows remain less clearly defined. This is because such assessments have mainly been emphasized within niche domains and still require broader community discussion and unify the standard. Due to the lack of widely accepted mainstream metrics for workflows, we follow the checklist of computational workflow FAIRification techniques (as shown in Figure 18) to address all relevant criteria. However, unlike dataset assessment, we do not assign detailed scores.

Based on the engineering design, we have addressed all the listed criteria, fully aligning with all 10 FAIRification recommendations. According to this standard, our computational workflow shows a high level of FAIRness.

9 Discussion

This study provides a practical solution for enhancing the FAIRness, interoperability and reusability of marine acoustic data for ecosystem research. In this chapter, we will discuss about the results we achieved, the strengths of the study, the limitations needed to be addressed in the future work.

9.1 Interpretation of the results

In this research, the research question is: How can the marine acoustic data pipeline and ontology be designed to address data consistency, support various ML tasks, and align with the FAIR principles for effective ecosystem research. To answer the question, our experiment approaches this question through four distinct aspects, which we discuss sequentially below.

9.1.1 Data integration and ontology design

Through a broad review and expert interviews, we found that the marine acoustic data sources are very federated with various data structures. So data was collected and first transformed into standard clips. For better data reuse, we also designed our ontology for this specific research. Because of the cross domain knowledge of marine, acoustic and machine learning, different domain ontologies were selected linked as well. Our ontology not only captures metadata and contextual information but also facilitates ML annotations. This supports downstream ML tasks by enabling consistent labeling and efficient data extraction across datasets. The FAIRification process guideline is the core to guide this design.

9.1.2 An optimized data framework

Through the literature review we identified three representative frameworks of soundscape ML applied in marine research. These frameworks served as key references from which we extracted a common data preprocessing pipeline, forming the basis for a standardized acoustic feature extraction workflow. Once raw audio was processed into unified vector representations, different ML tasks diverged from this point.

For soundscape level ML, our framework includes commonly used unsupervised learning algorithms such as PCA, UMAP, K-Means, and GMM. Research teams focused on abnormal acoustic events can directly utilize the GMM-based anomaly

detection outputs. These results are updated in the metadata file under the event label field, enabling downstream team fast querying for model training. This design facilitates cross-team reuse of both data and ML workflows, significantly enhancing research efficiency, interoperability and reusability.

9.1.3 Data Pipeline engineering and testing

Once the framework was designed, the actual engineering process became straightforward, allowing for modular implementation in Python. Standardized acoustic clips and metadata files were used as inputs for testing. The successful test results (including feature extraction, biodiversity assessment, and anomaly detection) demonstrated that the pipeline functions effectively and fulfilled its intended multi-functional purpose. Furthermore, by integrating Snakemake, the entire computational workflow was standardized, enabling pipeline level sharing and reuse.

9.1.4 FAIR maturity assessment

Finally, we conducted FAIR maturity assessments for both the dataset and the pipeline as key digital assets. The results indicated a high level of FAIR alignment, showing their readiness to be published and shared as contributions to open science.

Through the above four steps, we explored a standardized process for building a FAIR pipeline. Although this study specifically focuses on the application of soundscape level ML with marine acoustic data for ecosystem research, the standardized process has high transparency and has the potential to be adapted to other domains.

9.2 Strengths of the study

A key strength of this study is its successful integration of marine soundscape data into a ontology framework. By aligning metadata with established Semantic Web vocabularies and combining ontologies from multiple domains, including marine ecology, acoustics, and machine learning, the metadata design enhances machine readability and supports cross research data reuse. This provides an important foundation for federated data integration, making it especially suitable for AI-driven applications in marine ecology.

Additionally, our metadata design includes task-relevant labels to support result storage and data query for multiple machine learning tasks, such as biodiversity

assessment, anomaly detection, and species identification. Since the algorithms themselves are modular, the framework offers strong flexibility for integrating different new models in the future.

This multi-task compatibility overcomes the common limitation in ecological data pipelines, which are often designed for single purpose tasks. It also offers a new rethinking for cross-team collaboration by identifying shared preprocessing needs across different research groups. This approach reduces redundant data preparation, with the metadata file as a bridge, linking to the same preprocessed data while supporting various analytical goals, promoting interoperability and minimizing unnecessary duplication.

Finally, not only the FAIRified dataset but also the computational workflow are significantly contribute to open ecological science. By building a pipeline aligned with the FAIR principles, this study provides a replicable and transparent experience for future implementations.

9.3 Limitations

While it shows several strengths, there are also limitations to be discussed.

In this study, the main focus is on data FAIRification and data framework design. The calibration, as the specialized step in acoustic signal processing, is mentioned as a key procedure to reduce structural bias in data modeling caused by differences in devices and sampling methods. However, we do not compare or implement different calibration algorithms in detail.

Addtionally, from a data lifecycle perspective, unlike other original metadata, the labels used in 'edam:Machine learning' class are process-annotated, and therefore are not provenance features. In our research, we only describe such labels in the paper, but did not establish a more standardized annotation method to distinguish them from other metadata (emphasizing that they are assumed outcomes), nor did we implement the annotation to update or record whether these labels have been reassessed by researchers.

Also from the algorithmic perspective, some machine learning models - especially black-box types like neural networks, may lack of some explainability. To ensure

robustness and trust in the results, it is necessary to perform cross-validation and support model outputs with interpretable acoustic indices and external references such as historical literature or statistical records or even multi-modal images. These measures will help calibrate evaluation outcomes and strengthen the overall credibility and utility of the computational pipeline.

This pipeline method has so far only been tested on small-sample centralized-stored datasets, and its scalability in larger, real-world scenarios requires further exploration. In theory, acoustic data can be streamed in real time into the data pipeline for dynamic analysis; however, due to the lack of relevant hardware and experimental setup, real-time deployment remains untested and is an important direction for future implementation.

From the perspective of source-level data handling, PAM datasets, which are often large in volume, are typically stored on cloud platforms such as Google Cloud Platform (GCP). In this context, the entire computational pipeline can be executed within the cloud environment. Compared to traditional workflow management tools like Snakemake, cloud platforms offer a more accessible solution, especially for users with limited coding experience, as Snakemake requires deeper knowledge of command-line settings and scripting.

Alternatively, decentralized storage infrastructures such as FAIR Data Points (FDP) should also be considered. If participating institutions store acoustic datasets using a unified metadata structure, FDP enables seamless and efficient querying across distributed sources without copying data again, this would be a critical step toward enhancing data interoperability and reusability in future collaborative projects.

9.4 Future work

Regarding of the above limitation, more work is needed in future studies. Future work could therefore focus on three aspects: better calibration implement, better ML labels annotation standard and cross validation workflow, and further testing in a decentralized environment.

As future work, more detail calibration methods can be explored to solve the structural bias caused by different recording devices and sampling ways. Also, developing a more standardized annotation format for ML labels, clearly distinguishing them from

provenance metadata and tracking reassessment status, will improve transparency and reusability. Cross validation workflow could be another important module to be embedded in the data framework.

For the real-world use case scenario, improving scalability and exploring real-time streaming solutions is necessary, as well as leveraging cloud-based solutions, such as GCP, or decentralized infrastructures like FAIR Data Points for further testing.

10 Conclusion

This study aims to optimize a FAIR machine learning data pipeline by extending an ontology to enhance the interoperability and reusability of marine acoustic data for ecological research. Our pipeline is well realized for integrating data for cross-task use and provides new ideas for more collaborative data-driven marine research.

The following conclusions are drawn based on the research sub-objectives:

RO 1: Integrate Marine Soundscape Data with Optimized Ontology to Support Subsequent ML Tasks

This research successfully integrated marine soundscape data into an ontology-based framework, improving data interoperability and providing a standardized structure for future machine learning tasks. The optimized ontology enhances data reusability, ensuring that the data can be easily adapted for multiple ecological applications. By ensuring a semantically rich metadata schema, the integration supports seamless transitions between different ML tasks and use cases.

RO 2: Address Inefficiencies and Propose the Optimization Solution of the Data Pipeline to Support Diverse ML Tasks

The study successfully identified key inefficiencies in current marine acoustic data pipelines, such as limited support for diverse machine learning applications and insufficient FAIRification of the whole computational workflow. In response, the research proposed an optimized data pipeline that integrates federated data sources and applies standardized metadata through ontology extensions. Along with metadata file continually updated for data querying, it reduces data redundancy, ensures consistency, and enables more efficient reuse of the data for various ML tasks. Furthermore, the pipeline supports scalability, allowing it to accommodate research requirements and increasing data volumes.

RO 3: Set Up the Pipeline and Experiment to Test Its Performance, and Document It

An experimental setup was developed to test the performance of the optimized pipeline, focusing on its ability to handle various acoustic data and multiple machine learning tasks. The engineering result showed the pipeline's efficiency in processing integrated datasets, its flexibility in supporting various ecological research tasks, and its ability to data reuse. The entire workflow, from data integration to ML model

application, was documented in FAIR standard, providing a blueprint for future use in marine ecosystem open science.

RO 4: Use the FAIR Principle to Evaluate the Data Pipeline Maturity, Consistency, and Quality

Under the FAIR principles guidance, the data pipeline was evaluated of maturity, consistency, and overall data quality. The assessment revealed that the optimized pipeline significantly enhances the FAIRness of marine acoustic data, ensuring that it is not only accessible and reusable but also easily discoverable and interoperable across different research platforms. The pipeline's alignment with the FAIR principles supports its adoption for global ecological research collaborations.

Overall, through responding the above research sub-objectives, we conduct our research and bring some novelties. First is the marine acoustic data FAIRificaiton, building a cross domain ontology and federated data integration for marine acoustic data. Second is the consideration of multi-task ML requirements and cross team cooperation. When applying ML, we also consider downstream tasks, and the bridge - metadata file enabling efficient querying and filtering. Third is the FAIR computational workflow, following the FAIR workflow process, we also make the entire computational pipeline FAIR for marine ecosystem Open Science.

Although this study shows a practical solution to FAIRify marine acoustic data, more efforts should be made to improve the calibration procedure, and the robustness and interpretability of labels annotated by machine learning models by cross-validation with acoustic indices, historical references, multi-modal data or experts' adjustment. These directions are important for transforming the pipeline into a real-world, scalable, and trustworthy architecture for marine ecosystem research.

It has so far only been tested on small, centralized stored historical dataset. Future work should focus on real-time monitoring and decentralized storage solutions. Mature cloud infrastructures or decentralized data systems such as FAIR Data Points can further improve accessibility, interoperability, and collaborative efficiency.

In conclusion, this research provides an optimized solution to enhance the integration, consistency, and application of marine acoustic data. By optimizing the data pipeline and ensuring alignment with the FAIR principles, this study supports the growth of

marine ecosystem research and facilitates the broader use of data in machine learning applications for biodiversity conservation and ecological studies. It represents a first step toward building a more practical and interoperable foundation for marine acoustic data, supporting both ecological understanding and future machine learning applications through a FAIR and collaborative approach.

References

1. Gouvêa, T. S., Kath, H., Troshani, I., Lüers, B., Serafini, P. P., Campos, I. B., Afonso, A. S., Leandro, S., Swanepoel, L. H., Theron, N., Swemmer, A. M., & Sonntag, D. (2023). Interactive machine learning solutions for acoustic monitoring of animal wildlife in biosphere reserves. International Joint Conference on Artificial Intelligence. https://doi.org/10.24963/ijcai.2023/711

2. Tanhua, T., Tanhua, T., Pouliquen, S., Pouliquen, S., Hausman, J., Hausman, J., O'Brien, K. M., O'Brien, K., O'Brien, K., Bricher, P., Bricher, P., Bricher, P., Bricher, P., Bruin, T. D., Bruin, T. W. A. D., de Bruin, T., de Bruin, T., Buck, J., Buck, J., … Zhao, Z. (2019). Ocean FAIR data services. Frontiers in Marine Science. https://doi.org/10.3389/fmars.2019.00440

3. Lin, T., Lin, T.-H., Akamatsu, T., Akamatsu, T., Sinniger, F., Sinniger, F., Harii, S., & Harii, S. (2021). Exploring coral reef biodiversity via underwater soundscapes. Biological Conservation. https://doi.org/10.1016/j.biocon.2020.108901

4. Jones, D. O. B., Jones, D. B., Schoening, T., Durden, J. M., Durden, J. M., Faber, C., Faber, C., Felden, J., Felden, J., Heger, K., Heger, K., Hoving, H.-J. T., Hoving, H.-J. T., Kiko, R., Kiko, R., Köser, K., Köser, K., Krämmer, C., Krämmer, C., … Zurowietz, M. (2022). Making marine image data FAIR. Scientific Data. https://doi.org/10.1038/s41597-022-01491-3

5. Blumberg, K. L., Blumberg, K., Ponsero, A. J., Ponsero, A. J., Bomhoff, M., Bomhoff, M., Wood-Charlson, E. M., Wood-Charlson, E. M., DeLong, E. F., DeLong, E. F., Hurwitz, B. L., & Hurwitz, B. (2021). Ontology-enriched specifications enabling findable, accessible, interoperable, and reusable marine metagenomic datasets in cyberinfrastructure systems. Frontiers in Microbiology. https://doi.org/10.3389/fmicb.2021.765268

6. Jech, J. M., Zydlewski, G., Lebourges-Dhaussy, A., & Stevens, J. R. (2024). Ushering in a new era in fisheries and plankton acoustics. ICES Journal of Marine Science. https://doi.org/10.1093/icesjms/fsae112

7. Goble, C., Goble, C., Cohen-Boulakia, S., Cohen-Boulakia, S., Soiland-Reyes, S., Soiland-Reyes, S., Garijo, D., Garijo, D., Gil, Y., Gil, Y., Crusoe, M. R., Crusoe, M. R., Peters, K., Peters, K., Schober, D., & Schober, D. (2020). FAIR computational workflows. Data Intelligence. https://doi.org/10.1162/dint_a_00033

8. Woodall, L. C., Woodall, L. C., Talma, S., Talma, S., Steeds, O., Steeds, O., Stefanoudis, P. V., Stefanoudis, P. V., Jeremie-Muzungaile, M.-M., Jeremie-Muzungaile, M.-M., de Comarmond, A., & de Comarmond, A. (2021). Co-development, co-production and co-dissemination of scientific research: A case study to demonstrate mutual benefits. Biology Letters. https://doi.org/10.1098/rsbl.2020.0699

9. Wilkinson, M. D., Wilkinson, M., Wilkinson, M. D., Wilkinson, M. D., Dumontier, M., Dumontier, M., Aalbersberg, Ij. J., Aalbersberg, Ij. J., Appleton, G., Appleton, G., Axton, M., Axton, M., Axton, M., Axton, M., Baak, A., Baak, A., Blomberg, N., Blomberg, N., Boiten, J., … Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data. https://doi.org/10.1038/sdata.2016.18

10. Maureaud, A., Palacios-Abrantes, J., Kitchel, Z. J., Mannocci, L., Pinsky, M., Fredston, A., Beukhof, E., Forrest, D., Frelat, R., Palomares, M. L., Pécuchet, L., Thorson, J., Denderen, P. D. V., & Mérigot, B. (2024). FISHGLOB_data: An integrated dataset of fish biodiversity sampled with scientific bottom-trawl surveys. Scientific Data. https://doi.org/10.1038/s41597-023-02866-w

11. Jacobsen, A., Jacobsen, A., Kaliyaperumal, R., Kaliyaperumal, R., da Silva Santos, L. O. B., da Silva Santos, L. O. B., Mons, B., Mons, B., Schultes, E., Schultes, E. A., Roos, M., Roos, M., Thompson, M., Thompson, M., & Thompson, M. (2020). A generic workflow for the data fairification process. Data Intelligence. https://doi.org/10.1162/dint_a_00028

12. Wilkinson, S. R., Aloqalaa, M., Belhajjame, K., Crusoe, M., Kinoshita, B., Gadelha, L. M. R., Garijo, D., Gustafsson, O. J. R., Juty, N., Kanwal, S., Khan, F. Z., Köster, J., Gehlen, K. P., Pouchard, L., Rannow, R. K., Soiland-Reyes, S., Soranzo, N., Sufi, S., Sun, Z., … Goble, C. A. (2024). Applying the FAIR principles to computational workflows. Scientific Data. https://doi.org/10.1038/s41597-025-04451-9

13. Nieto-Mora, D. A., Rodríguez-Buriticá, S., Marín, P. A. R., Martínez-Vargas, J. D., & Isaza, C. (2023). Systematic review of machine learning methods applied to ecoacoustics and soundscape monitoring. Heliyon. https://doi.org/10.1016/j.heliyon.2023.e20275

14. Mahale, V., Chanda, K., Chakraborty, B., Salkar, T., & Sreekanth, G. B. (2023). Biodiversity assessment using passive acoustic recordings from off-reef location—Unsupervised learning to classify fish vocalization. Journal of the Acoustical Society of America. https://doi.org/10.1121/10.0017248

15. Buscaino, G., & Buscaino, G. (2022). Passive acoustics to study marine and freshwater ecosystems. Journal of Marine Science and Engineering. https://doi.org/10.3390/jmse10070994

16. Williams, B., Merrienboer, B. V., Dumoulin, V., Hamer, J., Triantafillou, E., Fleishman, A., McKown, M., Munger, J. E., Rice, A. N., Lillis, A., White, C. E., Hobbs, C. A. D., Razak, T. B., Jones, K. E., & Denton, T. (2024). Leveraging tropical reef, bird and unrelated sounds for superior transfer learning in marine bioacoustics. arXiv.Org. https://doi.org/10.48550/arxiv.2404.16436

17. Williams, B., Williams, B., Lamont, T. A. C., Lamont, T. A. C., Chapuis, L., Chapuis, L., Harding, H. R., Harding, H. R., May, E. B., May, E. B., Prasetya, M. E., Prasetya, M. E., Seraphim, M. J., Seraphim, M. J., Jompa, J., Jompa, J., Smith, D. J., Smith, D. J., Janetski, N., … Simpson, S. D. (2022). Enhancing automated analysis of marine soundscapes using ecoacoustic indices and machine learning. Ecological Indicators. https://doi.org/10.1016/j.ecolind.2022.108986

18. Roca, I., Roca, I. T., Roca, I. T., van Opzeeland, I., van Opzeeland, I., & van Opzeeland, I. (2019). Using acoustic metrics to characterize underwater acoustic biodiversity in the Southern Ocean. Remote Sensing in Ecology and Conservation. https://doi.org/10.1002/rse2.129

19. Elwen, S. H., Fearey, J., Ross-Marsh, E. C., Thompson, K. F., Maack, T., Webber, T., & Gridley, T. (2023). Cetacean diversity of the eastern South Atlantic Ocean and Vema Seamount detected during a visual and passive acoustic survey, 2019. Journal of the Marine Biological Association of the United Kingdom. https://doi.org/10.1017/s0025315423000255

20. Olsen, M. G., Halvorsen, K., Jiao, L., Knausgård, K. M., Martin, A. H., Moyano, M., Oomen, R. A., Rasmussen, J. H., Sørdalen, T. K., & Thorbjørnsen, S. H. (2021). Unlocking the potential of deep learning for marine ecology: Overview, applications, and outlook. arXiv.Org. https://doi.org/null

21. Domingos, L. C. F., Santos, P. E., Skelton, P. S. M., Brinkworth, R. S. A., & Sammut, K. (2022). An investigation of preprocessing filters and deep learning methods for vessel

type classification with underwater acoustic data. IEEE Access : Practical Innovations, Open Solutions. https://doi.org/10.1109/access.2022.3220265

22. Nolasco, I., Ghani, B., Singh, S., Vidaña-Vila, E., Whitehead, H., Grout, E., Emmerson, M., Jensen, F. H., Kiskin, I., Morford, J., Strandburg-Peshkin, A., Gill, L., Pamuła, H., Lostanlen, V., & Stowell, D. (2023). Few-shot bioacoustic event detection at the DCASE 2023 challenge. arXiv.Org. https://doi.org/10.48550/arxiv.2306.09223

23. Lou, R., Lou, R., Lv, Z., Lv, Z., Lv, Z., Dang, S., Dang, S., Dang, S., Su, T., Su, T., Li, X., & Li, X. (2021). Application of machine learning in ocean data. Multimedia Systems. https://doi.org/10.1007/s00530-020-00733-x

24. Sethi, S. S., Sethi, S. S., Jones, N., Jones, N. S., Fulcher, B. D., Fulcher, B. D., Picinali, L., Picinali, L., Clink, D. J., Clink, D. J., Klinck, H., Klinck, H., Orme, C. D. L., Orme, C. D. L., Wrege, P. H., Wrege, P. H., Ewers, R. M., & Ewers, R. M. (2020). Characterizing soundscapes across diverse ecosystems using a universal acoustic feature set. Proceedings of the National Academy of Sciences of the United States of America. https://doi.org/10.1073/pnas.2004702117

25. Williams, B., Belvanera, S. M., Sethi, S., Lamont, T. A. C., Jompa, J., Prasetya, M. E., Richardson, L., Chapuis, L., Weschke, E., Hoey, A., Beldade, R., Mills, S. C., Mills, S., Haguenauer, A., Zuberer, F., Simpson, S. D., Curnick, D., & Jones, K. E. (2024). Unlocking the soundscape of coral reefs with artificial intelligence: Pretrained networks and unsupervised learning win out. bioRxiv : The Preprint Server for Biology. https://doi.org/10.1101/2024.02.02.578582

26. Niu, H., Li, X., Zhang, Y., & Ji, X. (2023). Advances and applications of machine learning in underwater acoustics. Intelligent Marine Technology and Systems. https://doi.org/10.1007/s44295-023-00005-0

27. de Visser, C., Johansson, L., Kulkarni, P., Mei, H., Neerincx, P., van der Velde, K. J., Horvatovich, P., van Gool, A. J., Swertz, M. A., Hoen, P. A. C. 't, & Niehues, A. (2023). Ten quick tips for building FAIR workflows. Plos Computational Biology. https://doi.org/10.1371/journal.pcbi.1011369

28. https://medium.com/@lalit.k.pal/image-classification-a-comparison-of-dnn-cnn-and-transfer-learning-approach-704535beca25

29. Manna, G. L., Manna, G. L., Picciulin, M., Picciulin, M., Crobu, A., Crobu, A., Perretti, F., Perretti, F., Ronchetti, F., Ronchetti, F., Manghi, M., Manghi, M., Ruiu, A., Ruiu, A., Ceccherelli, G., & Ceccherelli, G. (2021). Marine soundscape and fish biophony of a Mediterranean marine protected area. PeerJ. https://doi.org/10.7717/peerj.12551

30. Pijanowski, B. C., Pijanowski, B. C., Villanueva-Rivera, L. J., Villanueva-Rivera, L. J., Dumyahn, S. L., Dumyahn, S. L., Farina, A., Farina, A., Farina, A., Krause, B., Krause, B., Napoletano, B. M., Napoletano, B. M., Napoletano, B. M., Gage, S. H., Gage, S. H., Pieretti, N., & Pieretti, N. (2011). Soundscape ecology: The science of sound in the landscape. BioScience. https://doi.org/10.1525/bio.2011.61.3.6

Appendix

Appendix A. literature review process

This is the full literature review process conducted in Chapter 3, in three main steps:

1) Conducted exploratory search in the beginning. Used the Google_Scholar (GS) database, with core keywords 'soundscape marine machine learning biodiversity'. Sorted out the relevant papers in the top20 search result as the seed papers. Expand more relevant articles for reference in the process of reading seed papers.

2) According to the systematic review method by D.A.Nieto et al.[13], we also selected ScienceDirect (SD) to carry out the systematic search because of the similar study fields. To define the search equation, Boolean filters provided by ScienceDirect were applied across three aspects: 1) terms in full articles; 2) terms in titles, abstracts, or keywords; 3) article types. The filter setting was performed according to Figure A1. Then used the PRISMA flow method to manage the whole review process. During the paper screening step, used AsReview tool to manually label all the papers into 'Relevant' or 'Irrelevant', according to the titles and abstracts. Sorted out all the 'Relevant' papers for further retrieval.

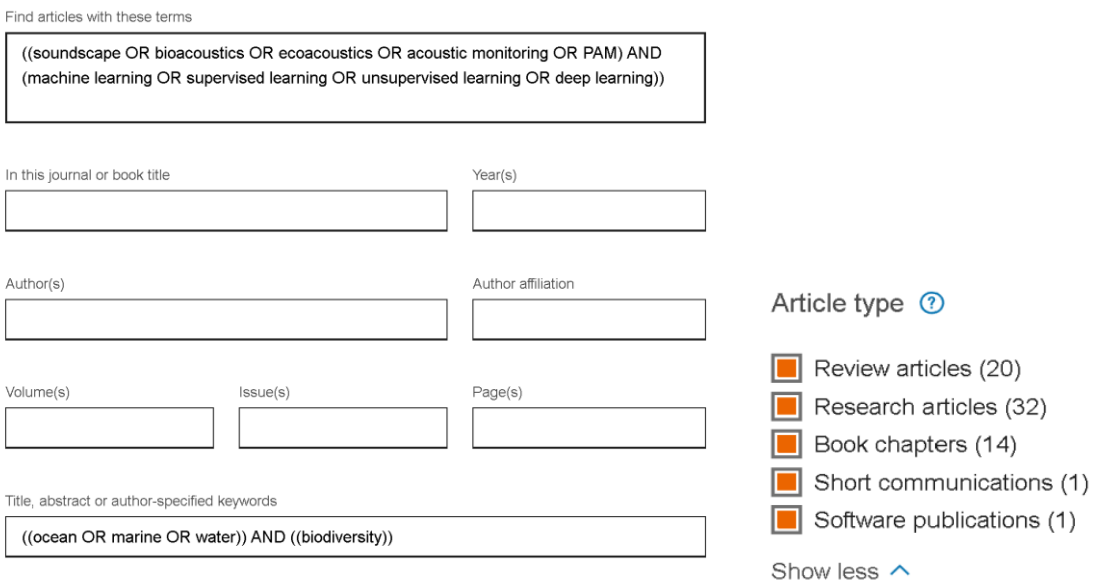

Figure A1: Filter query setting in ScienceDirect

3) Considering that there might be papers missing from the systematic search using only the ScienceDirect database, we also used ResearchRabbit (a database that searches for papers through the citation network, RR,

https://www.researchrabbit.ai/) for secondary supplementation. We put all GoogleScholar seed papers, extended reading papers and relevant papers screened in ScienceDirect as a batch, into ResearchRabbit to generate citation network top50 result, and then filtered papers manually in the same way with PRISMA and AsReview (https://asreview.nl/).

The last literature search is conducted in March 9th, 2025. Although a large number of search results were obtained through keyword combinations, the requirement to meet all three criteria including 'marine', 'soundscape' and 'machine learning', led to a limited number of relevant articles, with their content being relatively concentrated. There are a total of 21 relevant papers in the systematic review result in 3 sources as Figure A2:

1) Exploration search in GoogleScholar top20: 6 papers are relevant, and extend another related 10 papers mentioned in them.
2) Systematic search in ScienceDirect: 2 papers are relevant. (The manual screening result in AsReview is as Figure A3).
3) Connection search in ResearchRabbit in top50: 3 papers are relevant. (The paper connection graph is as Figure A4, and the screening result in AsReview is as Figure A5).

Most papers were excluded between the retrieval and eligibility steps because they did not apply machine learning, relying only on sound indices statistics methods. Additionally, some studies were excluded because they focused on individual species classification or animal behavior rather than soundscape level ecosystem assessment.

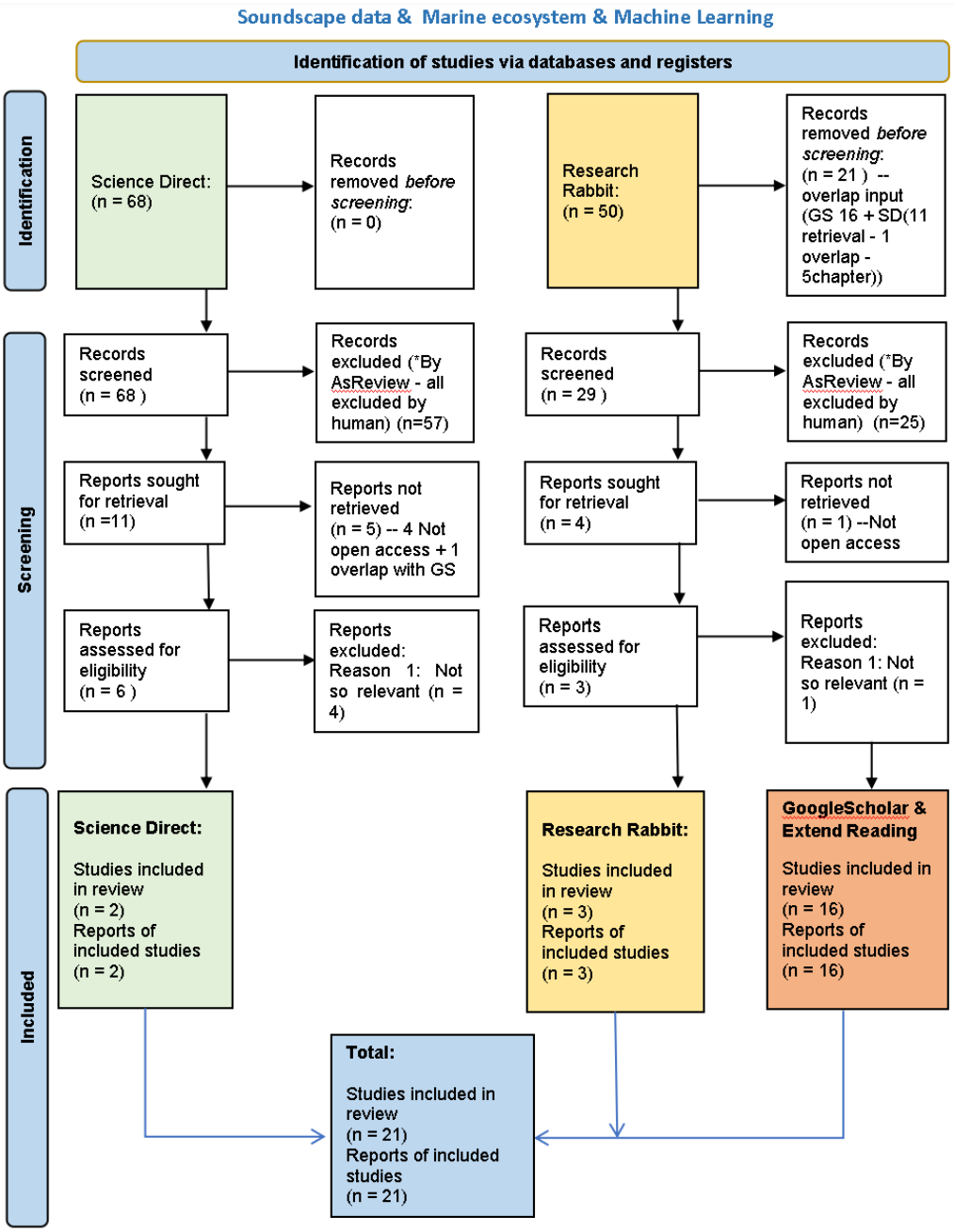

Figure A2: PRISMA 2020 flow diagram for systematic review

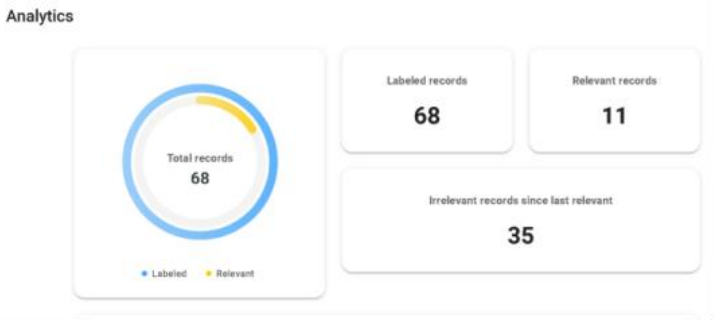

Figure A3:Screening of ScienceDirect result in AsReview

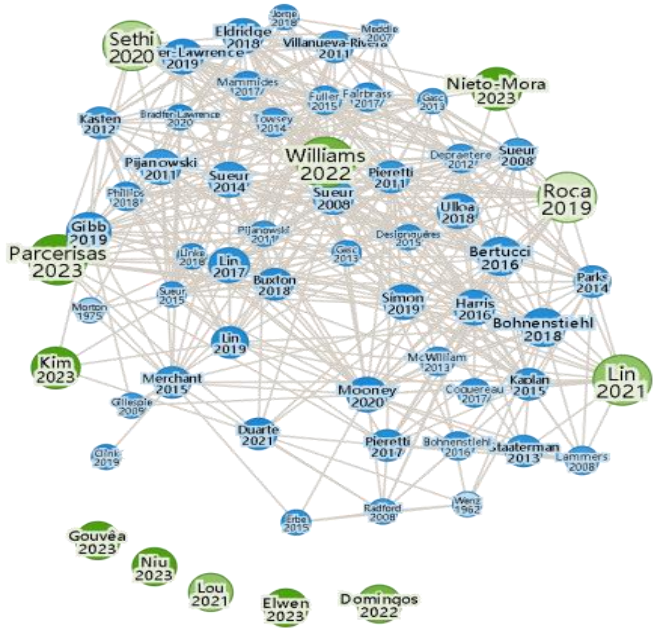

Figure A4: ResearchRabbit top50 paper network results

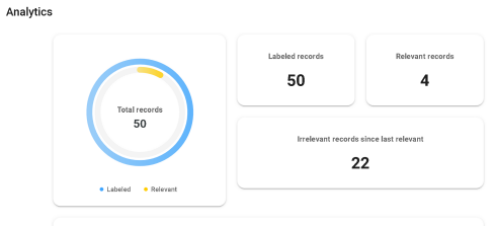

Figure A5:Screening of ResearchRabbit result in AsReview