genomic observatories

ASSEMBLE+
Association of European Marine Biological Laboratories Expanded

Data Management Plan (DMP)

for

Marine Biodiversity Observation Network for genetic monitoring of hard-bottom communities (ARMS-MBON)

Authors: Matthias Obst (University Gothenburg), Klaas Deneudt (VLIZ), Katrina Exter (VLIZ)

Version: LivingVersion

Date: 2020-05-07

Contents

1. Introduction

The European ARMS programme (ARMS-MBON) has established a network of Autonomous Reef Monitoring Structures (ARMS) placed in the vicinity of marine stations, ports, marinas, and Long-Term Ecological Research (LTER) sites distributed over Europe and polar regions. The aim of ARMS-MBON is to assess the status of, and changes in, hard-bottom communities of near-coast environments, using genetic methods supplemented with image analysis and visual inspection methods.

ARMS are passive monitoring systems originally developed during the Census of Marine Life project for the collection of marine fauna on and near the sea floor. Similarly to settlement plates, the ARMS units are stacks of plates that mimic the complex structure of the sea bottom. They are deployed on marine substrates and colonised by marine species, and after a period of time they are recovered by a team and taken apart to see who moved in. In the ARMS-MBON programme we have been deploying units since 2018 and which remain in place a few to many months.

One of our scientific goals is to identify newly arrived Non-Indigenous Species (NIS), facilitating an early warning system, and to track the migration of already known NIS in European continental waters. The data collected by the ARMS-MBON programme has the potential to be used for calculating Essential Biodiversity Variables (EBVs) on the distribution and abundance of benthic and non-indigenous species, as well as for continental-scale research on the community ecology and biogeography of benthic invertebrates.

The purpose of this data management plan (DMP) is to describe the data that will be generated by ARMS-MBON and to describe our plan to make these data findable, accessible, interoperable and re-usable (FAIR; Wilkinson et al., 2016). This DMP will cover the protocols and procedures for: constructing, deploying and collecting the ARMS units; processing the collected samples; and the molecular (genetic) and the image analysis. It will also describe the data flows, the data and metadata standards to be adopted, the archiving for long-term preservation, and the cataloguing to ensure the data are publicly accessible.

2. Data summary

2.1. Data description

The ARMS-MBON programme generates highly valuable ecosystem data collected through the deployment of ARMS units in European coastal waters and in the polar regions. ARMS are 3D passive sampling units attached to the sea floor and consisting of a stack of settlement plates. Deployments are set up close to ports, marinas, long-term ecological research sites, or places with high oceanographic connectivity.

The data collected by the ARMS programme include:

- **Administrative (meta)data**: general administrative data associated with the sampling activity. Examples are: the administrative documents to comply with the Nagoya Protocol and the Access and Benefit-sharing (ABS) requirements, the Material Transfer Agreement (MTA), and the IRCC codes or proof of "due diligence".

- **Technical (meta)data**:
  - **Molecular standard operating procedure (MSOP):** step-by-step instructions for carrying out the complex sampling, laboratory, and processing operations in a standardised and repeatable way
  - **Sample metadata:** detailed metadata on the where, when, who, and how for each of the samples collected
- **Genetic data**:
  - **Raw sequence data:** FASTQ files with raw sequence reads outputted by the sequencing platform
  - **Post-processing sequence data**: FASTA files with processed and quality controlled sequence data. Associated with these are the parameter files used by the software that processes the data
  - **Clustered/OTU table files:** processed sequence data clustered by Operational Taxonomic Units (OTU)
  - **Assigned OTU tables and/or ASV files**: OTU tables with taxonomic assignments and relative abundances
- **Image data**:
  - High resolution photographic pictures of settlement plates
  - High resolution photographic pictures of specimens for DNA barcoding
  - High resolution photographic pictures of habitat and sampling events
- **Occurrence records:**
  - Abundance and coverage of species directly observed by morphological identification
  - Abundance and coverage of species observed from image data
- **Environmental data**: measurements on the environment in which the ARMS was deployed. These data can be collected in-situ at the beginning, the end and/or throughout the deployment. Alternatively these data can be extracted or derived from other sources available for the deployment location.
- **Discovery metadata**: general informative description of a collected dataset, usually created in a data catalogue that makes the data publically findable.

2.2. Data collection and processing overview

The data collection is a distributed effort of the scientific institutes that are part of ARMS-MBON and that manage the deployments at the observatory stations. At the end of the deployment period, the ARMS units are retrieved and **sample processing** (Fig. 1[A]) is carried out in accordance with the official ARMS protocols (https://www.oceanarms.org/). Sample metadata, image data and species observation data, and any collected environmental data, are transferred to the **data management** platform, PlutoF, (Fig. 1[D]) by the station managers. Added to that are the required administrative data (MTA, ABS documents, etc.).

The DNA sequencing and **sequence processing** (Fig. 1[B]) is performed as a centralised effort coordinated by HCMR. The raw sequence data delivered by the sequencing lab as FASTQ files are uploaded to ENA (European Nucleotide Archive) and the ENA accession numbers are added to PlutoF.

As part of the sequence processing workflow, the raw sequence data are processed by PEMA (Zafeiropoulos et al., 2020), a metabarcoding analysis pipeline, producing OTU tables and ASV files with the taxonomies of the organisms found and their abundances in each sample. The generated FASTA files (including the input PEMA parameter files) are uploaded onto PlutoF (Fig. 1[E]) and may be also added to ENA.

The images taken of the ARMS plates will be analysed via **image processing** (Fig. 1[C]) software, producing CSV files containing the sample IDs, taxonomic assignments, and some measure of abundance/biomass. The code to use here is still under investigation (PhotoQuad is a contender, see Trygonis & Sini, 2012), and a set of recommendations for using whichever code is chosen will be written and provided to the consortium to follow.

Once a full year of ARMS-MBON (meta)data have been collected in PlutoF, they are copied to the Marine Data Archive (MDA) for long-term preservation, this including the: administrative metadata, technical (meta)data (MOPs and protocols, associated metadata), the genetic (meta)data (ENA run accession codes [cleaned and raw sequences], OTU tables and ASV files, associated metadata), the images and occurrence records and their metadata, and environmental (meta)data. These (meta)data are then gathered into formatted CSV and DarwinCore (OBIS-Event) files. These files, and the associated data and documents, are entered into the Integrated Marine Information System (IMIS) metadata catalogue, and that metadata record is made public. As data are subsequently analysed and additional species detections provided, these are added to the DwC files and the metadata record is updated to reflect this.

All sequence, image, and observation data have a moratorium period of one year from the point when the sequences (FASTQ) are uploaded to ENA and the run accession codes added to PluotF. Exceptionally, a period of two years will be imposed on the image data of 2019, due to delays caused by the Covid-19 event in 2020. The moratorium period is ensured for data in the ENA and MDA by having the data behind an account. After the year, all data are moved to be open access, with a CC BY licence. Note that the IMIS metadata record, and the CSV file containing the metadata and links to the actual data (Section 4), is always open access, so that the scope of the ARMS data can be accessed by anyone via the IMIS record.

PlutoF and IMIS both allow for the integration and quality control of the submitted data and can generate yearly ARMS datasets for **data archiving and publishing** (Fig. 1[E]): cleaned sequences on ENA as FASTA files, raw sequences on ENA as FASTQ files, and on EurOBIS and GBIF for the files with the collected parameters, sequence and image links, and taxonomy assignments, as Darwin Core Archive.

Where new versions of existing data or documents are archived and made public, version tracking will be done both in IMIS and the MDA.

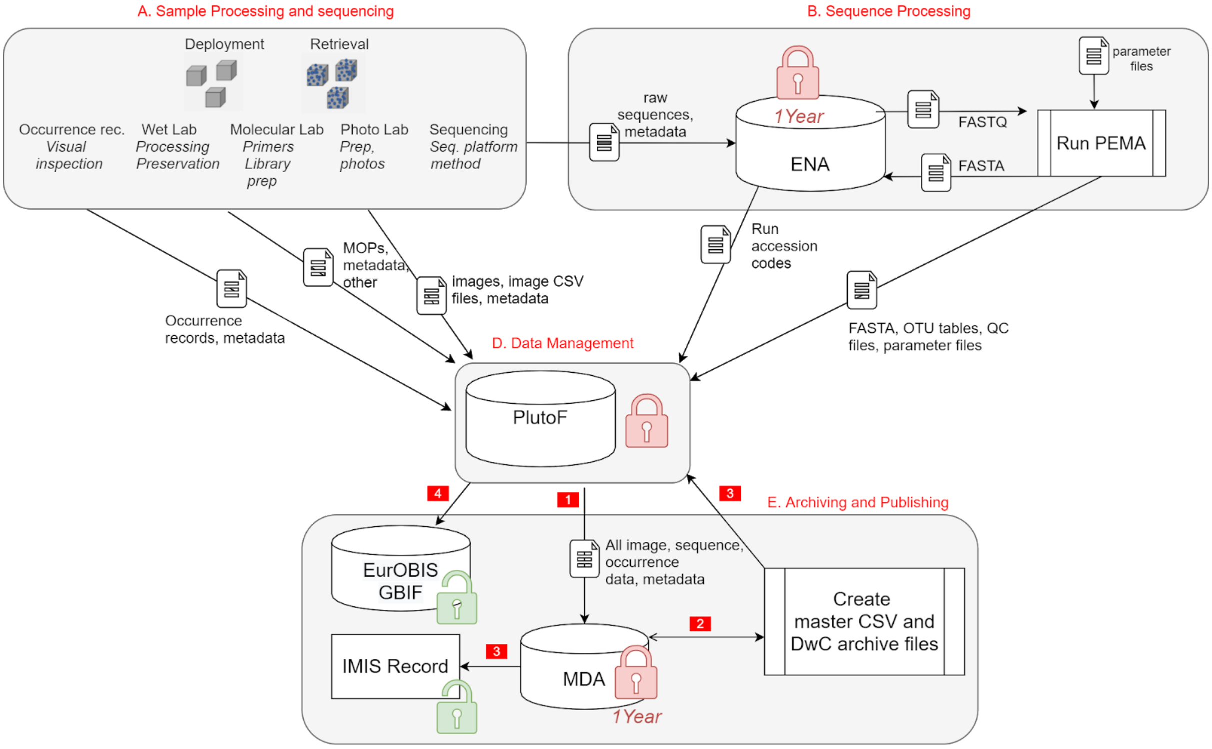

*Fig 1. ARMS data flow **A. Sample processing and sequencing** is organised by the ARMS-MBON network; **B. Sequence processing** archives raw sequences (FASTQ format) and generates cleaned sequences (FASTA files) as well as taxonomic assignments using PEMA; **C. Image processing** (still under development, not highlighted in the figure) will see the plate images analysed to produce taxonomic assignments and a measure of abundance/biomass; **D. Data management** is enabled by the PlutoF virtual research environment, that allows for online curation of all ARMS-MBON data by the consortium (access is limited to the consortium); **E. Data archiving and publication** will be done by storing the data in the MDA, creating metadata records in IMIS, and publishing to other data systems. PlutoF offers export functions for metadata, sequence data, and observational data. A one year moratorium period will be set for open access to the sequences, images, and occurrence records, after which the (meta)data will be published to EurOBIS and GBIF; the IMIS metadata record will always be public.*

2.3. Administrative (meta)data

Each partner in ARMS-MBON fills in a registration form from which administrative metadata are taken: contact details, observatory details, etc. A GDPR notice is included in this form.

Sampling permits are required for each partner. To comply with the Nagoya Protocol on Access and Benefit Sharing (ABS), it is necessary for each partner to request the necessary permits from their ABS National Focal Point. This is documented in our ABS HowTo. The resulting IRCC codes or proof of “due diligence” will be uploaded to PlutoF.

2.4. Technical (meta)data

The following technical documentation form the technical data:

- Molecular standard operating procedures (MSOPs) and Standard Operating Procedures (SOPs; which are included in the Handbook).
- Description of software processing: the sequence cleaning and species assignments done by PEMA, and the recommendations for running the chosen image analysis software

The technical metadata include the following:

- Observatory coordinates
- Dates of deployment and retrieval
- Sample fraction, filter, and for the 2018 data also the preservative
- Standardised IDs (following OBIS recommendations)

2.5. Genetic (meta)data

The following genetic data are collected:

- Raw sequences, stored in ENA as FASTQ files
- Cleaned sequences, as FASTA files
- OTU tables or ASV files and some QC output files produced by the PEMA processing of the FASTQ files, and the parameter files input into PEMA

The following genetic metadata are collected:

- ENA run accession codes for the FASTQ files (and FASTA files, where they are also archived). The raw sequence files (FASTQ files) are submitted using the ENA checklist. For each sample, metadata were added to the "sample_description": the collection year, coordinates and geographic location (country and/or locality), the fraction, the ARMS project, the preservation, and the amplified gene/region (e.g. "Sample from the motile fraction (500 um) of the ARMS plates in AZFP2 preserved in EtOH and amplified for the 18S rRNA"). All samples were tagged as "environmental_sample" and were submitted under the NCBI taxonomy ID 408172 corresponding to "marine metagenome".
- Details on the quality assessment, quantification and total volume of the DNA samples are stored in PlutoF.

2.6. Image (meta)data

The Handbook explains the protocols for taking images of the ARMS plates, specimens, the sampled habitat, and lists the metadata that should accompany the images. Which image-processing codes(s) to use on these images is still being investigated (PhotoQuad is a contender); once chosen, recommendations for processing with the code(s) will be added to PluotF and archived in the MDA. The output format, from which code is used, is recommended to be a CSV file with taxonomic assignments (with WoRMS Aphia IDs) and annotated biomass/abundance values, together with the sampleIDs and plateIDs (see Section 4). These files will be archived in PlutoF and the MDA, and the taxonomic assignments will be added to the DwC-formatted files provided via IMIS.

Any images that are annotated by the scientists will include some metadata such as: data creator, taxonomic assignments, and which plate the observations were made from.

2.7. Occurrence records (meta)data

Visual assessments (i.e. manual observations) of the ARMS plates can be made before they are processed into samples, for example while being retrieved or while being photographed, resulting in an "occurrence record". A template CSV file to hold the collected information can be found in the Handbook; this includes taxonomic assignments with WoRMS aphia IDs and material and image sample IDs.

2.8. Environmental (meta)data

At present no environmental metadata are collected for the ARMS-MBON data of 2018 and 2019. A set of recommendations covering what to collect and standards to use when recording the measurements will be created for the 2020 collection.

2.9. Discovery metadata

Discovery metadata are those that are given in a catalogue entry, a metadata record, for a dataset. As IMIS is the primary catalogue used by ARMS-MBOL, the discovery metadata are stored in the IMIS database and include the fields listed in Section 3.1; the public can search on any of these fields using the IMIS datasets search form (e.g. here); the metadata can be obtained in HTML, XML, EML, and JSON format.

3. Making data FAIR

The FAIR principles are about making data **Findable**, **Accessible**, **Interoperable**, and **Re-usable**. These principles are explored in Wilkinson et al. (2016), and in the main they are concerned about the machine-to-machine actionability of data, e.g. that data can be found not only by a human using a web-interface, but by a human-launched machine search "over the web". This requires that the data are F, A, I, and R to and for machines, that the data can be understood by a computer, not only a human.

Making data FAIR requires specific actions or arrangements to be included in the data management process. Here we describe those actions we will undertake.

3.1. Making data Findable

To make the ARMS data findable, to inform external users of the existence and scope of the collected data, discovery metadata are needed for all collected datasets. This is commonly realised by creating discovery metadata records for the collected data in a relevant online data catalogue. ARMS-MBON will use the Integrated Marine Information System (IMIS) catalogue as the primary metadata record for all its data. IMIS can be searched by humans via a web-interface, and machine-to-machine (m2m) searches can be performed via its various webservices. The human-readable metadata record is in HTML; they are also provided in XML, JSON, and EML metadata formats[1].

An overall parent metadata record for the ARMS-MBON programme will be constructed, with child records holding the (meta)data for each year of the programme. Some of the ARMS data of 2018 – a preliminary year during which the protocols and practises were still being investigated – are already catalogued in IMIS.

The following fields are created for the ARMS records in IMIS:

- Title, abstract, and description which covers the content of the dataset
- Name and contact details for the person responsible for the dataset
- A citation for the record
- Links to access the associated datasets (see Section 4 for more on what these associated datasets are)
- Licence of use
- Keywords chosen for the dataset
- A listing of the parameters measured
- The spatial, temporal & taxonomic coverage of the dataset
- A listing of people who contributed to the dataset
- Links to related and child/parent datasets

Upon creation, IMIS metadata records receive URIs based on their local IMIS identifier: these URIs are the tail part of the permanent and unique URLs each record receives.

DOIs for the IMIS metadata records can also be created by IMIS via collaboration with DataCite.

The (meta)data for each year of ARMS-MBON will be downloaded in bulk from PlutoF to the MDA. Automatic processes will be developed at VLIZ to organise those (meta)data into the chosen formats with the chosen standards, before archiving them in the MDA and creating the IMIS metadata record.

3.2. Making data Accessible

To make the ARMS data widely finadable and accessible, the (meta)data will be published from PlutoF towards a number of data infrastructures for re-distribution (see Fig. 1[E]). All of these (EurOBIS, GBIF) provide m2m access to data using standardised communications protocols.

ARMS-MBON data archived in the MDA for access via the IMIS metadata records, can be made open access or can limit access to permitted users[2], thus ensuring management of the chosen moratorium periods. Data in EurOBIS can have a moratorium period, however the aim is for open and freely-accessible data. GBIF requires the use of CC0, CC BY, or CC BY-NC licences: hence ARMS-MBOL data will be pushed to these external data infrastructures only after the moratorium period has elapsed.

Metadata and data in the chosen data infrastructures are never deleted, hence they are always accessible.

[1] The EML 2.2 metadata schema, which supports semantic annotation of metadata, will be added to IMIS. JSON-LD is also under consideration.

[2] The MDA's AAI interface is currently not m2m actionable, but will be in the future.

3.3. Making data Interoperable

Metadata are made interoperably by adhering to globally-accepted metadata standards. The IMIS datasets catalogue uses the EML metadata schema[3] (a standard for ecological data) as well as JSON and XML. GBIF also uses EML also, and EurOBIS discovery metadata are accessed via IMIS. IMIS uses a number of standards in its fields, in particular the World Register of Marine Species (WoRMS) for taxonomy; Marine Regions and FAO ASFA geoterms for geographic locations and for thematic keywords.

ARMS-MBOL data will be made interoperable by conforming to globally-accepted biodiversity data standards, primarily Darwin Core Archive, while publishing the data output towards integrative systems such as ENA and EurOBIS (see data archiving and publishing, Fig. 1[E]). WoRMS will be used as a standard vocabulary for resolving synonymy and assigning valid species names. The WoRMS aphia IDs will be included as a unique identifier for all identified taxa. Parameter names will be linked to the controlled vocabularies of the British Oceanographic Data Centre (BODC).

See Section 4 for more detail on the data formats we will adopt.

3.4. Making data Re-usable

IMIS metadata records will be created as soon as each year's worth of ARMS-MBON data have been gathered (i.e. images, raw sequences, and associated metadata and documentation). These records will be public, and the data file (CSV) with an overview of the data available for that year will be open access (CC BY). A citation for referencing the dataset is always included in the discovery metadata record. However, all data (images and raw sequences, and later then cleaned sequences and species identifications) will have a moratorium period of one year from when the raw sequences are placed in PlutoF; this allows the partners time to quality control, integrate and analyse the data and have priority on publishing the first scientific results. Exceptionally, a period of two years will be imposed on the image data of 2019, due to delays caused by the Covid-19 event in 2020.

Quality control of the datasets are done before uploading to the MDA, and any departures from the operating protocols are added as notes in PlutoF. Additionally, QC is done upon integration in data infrastructures such as Genbank and EurOBIS. This ensures that data are re-usable and that all required technical metadata for appropriate interpretation of the data is included.

Data will remain re-usable by archiving the data in MDA as a long-term repository.

4. The public data

Here we describe the formats of the (meta)data that are linked to the metadata record in IMIS, and which are exported to EurOBIS and GBIF via PlutoF.

[3] Currently EML 2.1.1. EML 2.2, which supports semantic annotation of metadata and hence is more interoperable than previous versions, will be added in the near future. Adding JSON-LD is also under consideration.

4.1. Metadata and data files

The following data are stored in PlutoF for each year of ARMS-MBON sampling:

1. ENA run accession codes for the raw sequences
2. Plate images and the image metadata CSV file (to link the images to the plate number they were taken from)
3. Material sample metadata
4. Observatory metadata
5. Cleaned sequences (FASTA) files
6. The parameter files used to run PEMA on the raw sequences
7. The OTU tables, ASV files, and the QC files produced by PEMA, for each processing carried out

Additionally, individual partners may add the following data to be also downloaded

8. Any other documentation required to understand the data
9. Occurrence records created from manual inspection of the ARMS units or the images, or as created from the cleaned sequences via their own processing

These data are downloaded to the MDA once a full year has been archived in PlutoF. This is likely to be carried out once the datasets 1-4 and 8-9 are ready, and then later when the datasets 5-7 have been created.

From this collection of (meta)data, CSV and Darwin Core files are created to link them together. These files are archived in the MDA and linked to the IMIS metadata record. The DwC files will be exported to PlutoF once they are complete, and from there will be exported to EurOBIS and GBIF.

4.2. Data formats

The CSV file format is provided for the “traditional” user of such data. An example of this file can be found attached to the IMIS metadata record for ARMS 2018. The purpose of this file is to list the geotemporal data for each ARMS sample from each ARMS observatory, and to include the run accession codes and links to the image files. These latter data will not be open access until the moratorium period ends, this being indicated by a column “open access/not open access” in the file. Any collected environmental values will also be included in this CSV file.

The DwC file will have the OBIS-Event format (for the data and the included metadata). A first file will be created with the data listed in steps 1-4, hence the occurrences part will be left blank. Once species identifications have been made via images, visual observation, or the sequences, these will be added to create full OBIS-Event format files with occurrence information. At the time of writing we have not created any DwC files, but they will be described here when that has happened. It is intended to follow all OBIS recommendations for this format of data.

5. Ethical & GDPR aspects

The ARMS network will take measures to be compliant with the EU regulations regarding the protection of personal data (http://ec.europa.eu/justice/data-protection/).

6. Responsibilities and security

Implementation of the DMP is necessary at both the central and local level. The final responsibility for implementing the DMP lies with the ARMS-MBON network. The partners are responsible for ensuring the DMP is being implemented at the local level. There will be central efforts to support the partners by organising the necessary training and guidelines.

7. References

Callahan, B. J., McMurdie, P. J., Rosen, M. J., Han, A. W., Johnson, A. J. A., & Holmes, S. P. (2016). DADA2: high-resolution sample inference from Illumina amplicon data. *Nature Methods*, *13*(7), 581.

Costello, M. J., Bouchet, P., Boxshall, G., Fauchald, K., Gordon, D., Hoeksema, B. W., Poore, G. C. B., van Soest, R. W. M., Stöhr, S., & Walter, T. C. (2013). Global coordination and standardisation in marine biodiversity through the World Register of Marine Species (WoRMS) and related databases. *PloS One*, *8*(1).

Hall, T., Biosciences, I., & Carlsbad, C. (2011). BioEdit: an important software for molecular biology. *GERF Bull Biosci*, *2*(1), 60–61.

Jain, M., Olsen, H. E., Paten, B., & Akeson, M. (2016). The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biology*, *17*(1), 239.

Kearse, M., Moir, R., Wilson, A., Stones-Havas, S., Cheung, M., Sturrock, S., Buxton, S., Cooper, A., Markowitz, S., & Duran, C. (2012). Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. *Bioinformatics*, *28*(12), 1647–1649.

Kumar, S., Tamura, K., & Nei, M. (1994). MEGA: molecular evolutionary genetics analysis software for microcomputers. *Bioinformatics*, *10*(2), 189–191.

McCarthy, C. (1996). Chromas version 1.45. *School of Health Science, Griffifth University, Gold Coast Campus, Queensland, Australia*.

Patterson, J., Chamberlain, B., & Thayer, D. (2004). Finch TV Version 1.4. 0. *Publ. by Authors*.

Ratnasingham, S., & Hebert, P. D. N. (2007). BOLD: The Barcode of Life Data System (http://www. barcodinglife. org). *Molecular Ecology Notes*, *7*(3), 355–364.

Trygonis, V., Sini, M., 2012. photoQuad: a dedicated seabed image processing software, and a comparative error analysis of four photoquadrat methods. *Journal of Experimental Marine Biology and Ecology*, 424–425, 99–108. doi:10.1016/j.jembe.2012.04.018

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., Blomberg, N., Boiten, J.-W., da Silva Santos, L. B., & Bourne, P. E. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, *3*.

Zafeiropoulos, H, Quoc, V. H., Vasileiadou, K., Potirakis, A., Arvantidis, C., Topalis, P., Pavloudi, C., Pafilis, E. (2020). PEMA: a flexible Pipeline for Environmental DNA Metabarcoding Analysis of the 16S/18S ribosomal RNA, ITS, and COI marker genes. *GigaScience*, Volume 9, Issue 3, giaa022.