Data Management Plan (DMP)

for

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

Authors: Katrina Exter (VLIZ), Matthias Obst (University Gothenburg), Christina Pavloudi (HCMR/George Washington University), Ioulia Santi (EMBRC)

Contact: katrina.exter@vliz.be

Version: Version 2

Date: 2023-02-15

Contents

Who and what are we?

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.

ARMS-MBON is a project that ran under ASSEMBLE Plus[1], from 2018 until the end of 2022. Thereafter ARMS-MBON merged with EMBRC's EMO BON project and therefore ARMS-BMON will run along two tracks:

1. The partners who are also part of EMBRC will continue their ARMS activities with the same sampling protocols as before. Some details of the policies, data management, and data collection have changed to accommodate this merger with EMO BON. The ARMS Handbook, SOPs, other documentation, and data entry templates can be found on the ARMS-MBON GitHub space, with the landing page https://data.arms-mbon.org/ giving easy entry to this space. Changes in this DMP for this merger will be pointed out in the text here, and it is expected that in 2024 the EMO BON and ARMS-MBON DMPs will be combined into one.
2. For those who are not in EMO BON but want to continue to sample with ARMS units on their own resources: it has not yet been decided what to do here, however it is most likely that these people will use the same templates, handbook, SOPs, and DMP as used by those participating in EMO BON.

[1]European Union's Horizon 2020 research and innovation programme under grant agreement No 730984

1. Registration and first steps

During the ASSEMBLE Plus years, each partner in ARMS-MBON filled in a registration form from which administrative metadata were taken: contact details, observatory details, etc. Observatory and ARMS unit names were also chosen at this point. This information was then transferred to the ARMS overview google sheet and to the starting pages for that partner in the data management platform we still use, PlutoF. *With the merger with EMO BON, this registration form is no longer to be used*, instead "an expression of interest" form will be created: when this is ready, we will provide its URL here. Meanwhile, for those who wish to express an interest in doing ARMS-MBON sampling, please contact Matthias Obst (matthias.obst@marine.gu.se>) who can tell you want to do.

Before any sampling activity, i.e. before any ARMS units are emplaced, it is necessary to decide how and what sort of sampling to perform. All of the necessary information can be found in the Handbook. Once these were decided, under ASSEMBLE Plus each partner had to then obtain the necessary local, national, and international sampling permits, including those related to the Nagoya Protocol on Access and Benefit Sharing (ABS; this is documented in the ARMS-MBON ABS HowTo guide). Under EMO BON, these permit steps are dealt with centrally and any new partner will be guided through the information they need to provide by the EMO BON secretariat.

2. ARMS-MBON data

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.

ARMS-MBON 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.

- From 2018 to 2022, i.e. before joining EMO BON, the sampling (meta)data were recorded in the data management platform, PlutoF, by the station managers, following instructions and templates provided by the data management team and as explained in the Handbook. DNA data that were produced by the sequencing team from the material samples shipped to HCMR, Crete, were archived in the European Nucleotide Archive (ENA) and the accession numbers were added to an ARMS overview google sheet.
- Under EMO BON, we are still using PlutoF to record some of the event metadata and in particular the images of the ARMS plates taken during each event and spreadsheets containing manual observations. Event and sample metadata are also required to be entered in logsheets, which are individual google sheets (per observatory and per sampling time) created by the EMO BON secretariat. The sequences produced from the material samples shipped to the EMO BON HQ in Paris will also be archived in ENA, and the accession numbers will be provided via a method yet to be decided.

The ARMS-MBON (meta)data are harvested into the ARMS GitHub site (https://github.com/arms-mbon for the GitHub-savvy, or this landing page otherwise) to be annotated, linked, and made FAIR and publicly-accessible. Now we have joined with EMO

BON, these data will instead be uploaded to the EMO BON GitHub site (see here for the Github space, we will add a link to a landing page when that exists).

Subsequent sequence analysis or image analysis outputs will also be made FAIR via the two above-mentioned GitHub sites. However, for the present we do not discuss such data in this DMP as these steps have not yet been completed.

2.1. The (meta)data we collect

The (meta)data collected by the ARMS-MBON programme include:

- **Administrative metadata**: general administrative data associated with the sampling activity, such as observatory and ARMS unit IDs, contact details, etc.
- **Procedural and sampling (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. Note those used during the ASSEMBLE Plus and the EMO BON periods are different.
    - **Sampling SOPs**: we follow the sampling SOPs of the Global ARMS program with some minor modifications.
    - **Sample/event metadata:** detailed metadata on the where, when, who, and how for each of the samples collected.
- **Genetic (meta)data**:
    - **Raw sequence data:** FASTQ files with raw sequence reads outputted by the sequencing platform.
    - **Metadata** associated with the production of these sequences.
- **Image data**:
    - High resolution **photographic pictures** of settlement plates: mandatory
    - Optional: high resolution photographic pictures of specimens for DNA barcoding, of the habitat and of sampling events
    - **Spreadsheets** providing image metadata, such as ARMS plate numbers, subject of the photograph, etc.
- **Occurrence records:**
    - **Spreadsheets of abundance and coverage of species** directly observed by morphological identification by eye.
    - Spreadsheets of abundance and coverage of species observed from the photographic images.
- **Discovery metadata**: general informative description of a collected dataset, usually added to a metadata record in a data catalogue that makes the data publically findable
- **Permit (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”. Under EMO BON these permits are dealt

with centrally and the metadata therefore also, and these will be made available at a suitable point. Before EMO BON it was required that each station dealt with these themselves, and kept the information and the permits locally to be provided as and when required. Therefore, there is no central store of permit (meta)data for the years 2018-2022.

No environmental data are collected by ARMS-MBON, due to the difficulty of making measurements over the month-long periods during which the ARMS units are submerged.

2.2. Data management steps

(1) The first steps of the data management are those related to the sampling events carried out by the sampling teams, following instructions in the Handbook. All observatory, sampling and event (meta)data are added to *PlutoF* by each partner, and some of the same metadata are also added to the ARMS overview google sheet. Once units have been retrieved, additional event and sample metadata are added to both these platforms. *Under EMO BON*, the ARMS overview google sheet is no longer used, instead the same metadata are added to each observatory's ARMS logsheets, which will be provided for each observatory when they join EMO BON.

(2) Once sufficient material samples have been shipped and received, they are processed, the FASTQ files uploaded to ENA (*European Nucleotide Archive*), and the ENA accession numbers are made available to the ARMS//EMO BON project members. These sequences are available to be used only by the project members for a period of 1 year initially, reduced to 6 months under EMO BON, and then they become open access. During the initial, ASSEMBLE Plus period of ARMS-MBON, the sequencing centre was HCMR who followed a MSOP, did the ENA uploading, and added the accession numbers to the ARMS overview google sheet. During the current, EMO BON period, the sequencing is done by Genoscope following a different MSOP and the EMO BON secretariat will do the uploading to ENA; the accession numbers will then be made available via a method yet to be determined.

(3) Once a sufficient number of events have been completed and/or sequences produced, the (meta)data from the various locations are harvested into the ARMS-MBON and/or EMO BON GitHub spaces. Normally this will be 2–4 times a year. We harvest from

- PlutoF for observatory, event, sampling, and material sample metadata, and the links to the image files and the spreadsheets of manual observations (for both the ASSEMBLE Plus and EMO BON periods)
- ARMS overview google sheet for the observatory, and some event, sampling, and material sample metadata and the ENA accession numbers (for the ASSEMBLE Plus period). Semantic annotations for all the metadata are also found in this google sheet
- EMO BON logsheets for observatory, event, sampling, and material sample metadata (for the EMO BON period)
- EMO BON spreadsheets for the ENA accession numbers (for the EMO BON period; still to be confirmed)

The ARMS-MBON GitHub space can be accessed from its landing page. The EMO BON GitHub space does not yet have a landing page because its space is still under construction, but it can be accessed from https://github.com/emo-bon.

At present, the images and manual observations that are uploaded to PlutoF are not copied into GitHub, rather their URLs are and we ensure that at the appropriate time, these data are openly accessible from those PlutoF URLs.

(4) All metadata undergo a quality control which are essentially sanity checks: are the same metadata from different sources the same, are all IDs correct, is the formatting correct, have all mandatory metadata been provided, etc. The images are annotated with metadata where those are provided (primarily the plate face and number). Where mistakes/missing metadata are identified, the observatories concerned are asked to make corrections/add information to PlutoF, the ARMS overview google sheet, or their EMO BON logsheets, and a new harvest is triggered, followed by another QC, etc until the QC is fully passed.

(5) Next, the metadata from the different sources are combined into a set of spreadsheets that allow for an easier browsing of the metadata, data (image and DNA), and events to-date. These can also be found in the ARMS-MBON and/or EMO BON GitHub spaces.

(6) Semantic descriptions of all the metadata are to be added to these combined data. This step is still underway, more information will be given once this step has been completed. This will allow for semantic uplifting, and the formatting of the data into formats such as turtle and JSON-LD.

(7) The (meta)data are formatted for different users, as and when requested/approved. Currently the ARMS-MBON data are provided for the LifeWatch Tesseract, which is a virtual research environment for the analysis of data on non-native and invasive species. The ARMS-MBON data are also linked to VLIZ's datasets catalogue, IMIS, (http://www.assembleplus.eu/information-system?module=dataset&dasid=8135) that will continuously be updated. The intention is to also publish the data on EurOBIS and/or GBif, but this step is still under investigation.

These steps are outlined in Fig. 1.

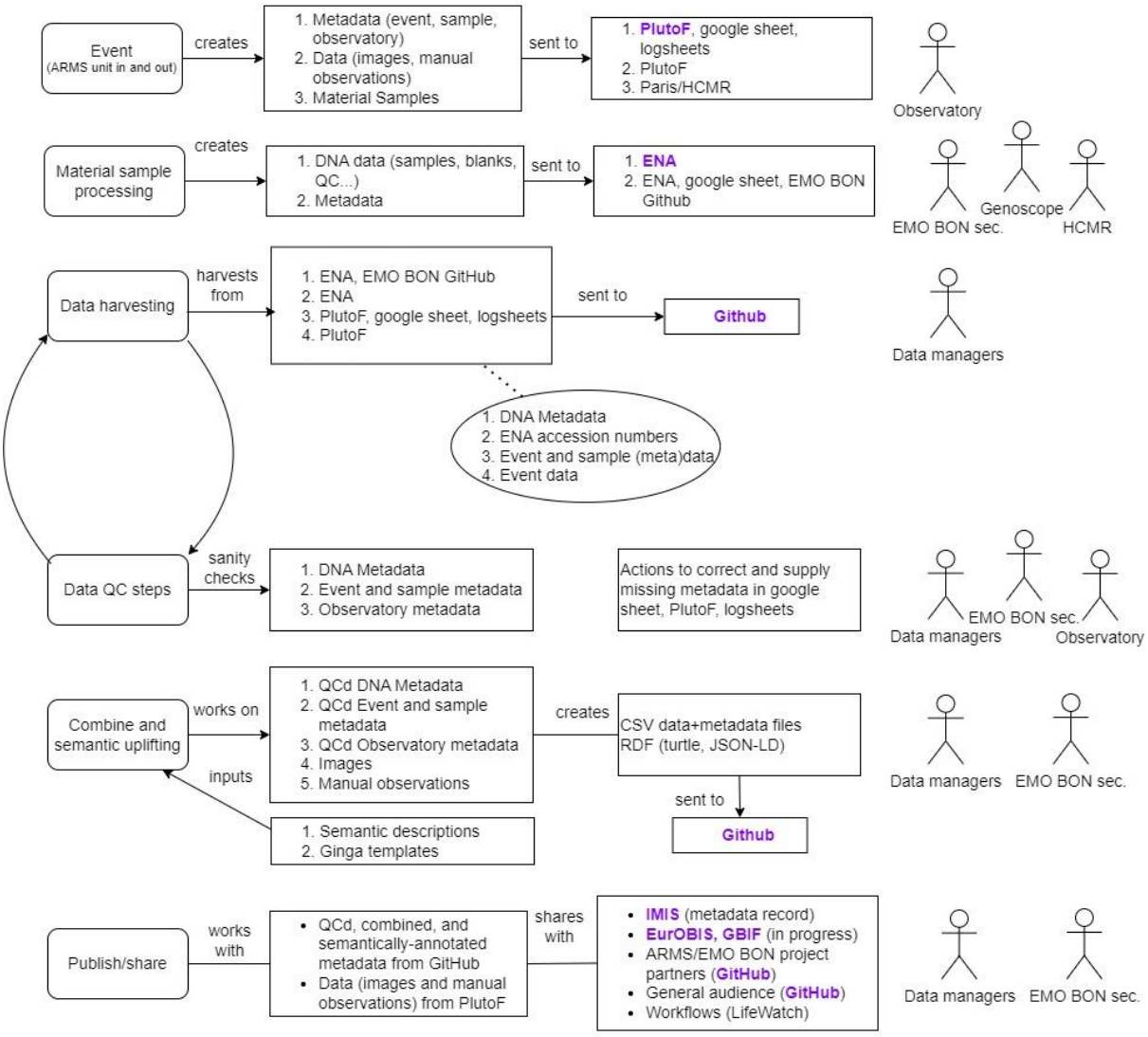

Fig 1. *Steps in the ARMS-MBOM work, from the sampling events to publishing/sharing data. Note included are steps related to image processing as this is yet to be established. Highlighted in colour are the external data archives/data spaces that we use, noting that publishing ARMS data on EurOBIS and GBIF is still a work in progress.*

2.3. Licences

All sequence, image, and observation data have a moratorium period of one year from the point when the sequences are uploaded to ENA, and then they are given the licence CC BY. Under EMO BON, this period is reduced to 6 months. 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 ENA by having the data behind an account, and likewise for the images and manual observations stored in PlutoF. Note that the metadata (in particular those in GitHub) are always open access, CC BY.

3. Data management platforms and data portals used

The data management platforms/online spaces that we use are highlighted in Fig. 1. Here we describe what we do with them in more detail.

3.1. PlutoF: observatory, event, sample (meta)data

PlutoF is a data management platform used by the project. In the ARMS-MBON space, each observatory is defined as a site, each ARMS unit is described as a sub-site linked to the parent site, and under those sub-sites the sampling events are recorded and photographs and other files are uploaded. The observatories are responsible for adding their administrative and sampling/event metadata, following instructions provided in the Handbook.

- **IDs** Identifiers
    - For the observatories and ARMS units in each observatory. The ARMS IDs belong to the location rather than the physical unit itself, as the units are taken up, cleaned, and re-used over and again. A new unit in an old location will receive the old ARMS ID.
    - Material sample IDs after the samples are shipped (adding _[fraction][filter size]_[preservative] to the event ID e.g. MF500_DMSO). These are also to be written on the samples as they are shipped to HCMR/Paris and/or biobanked locally, hence the actual *material* sample ID is supposed to be the same as the *digital* sample ID
    - Sampling event IDs, following a template that was: ARMS_[Observatory ID]_[ARMS ID]_[dateIn]_[dateout] with dates given as YYYYMMDD, and is now: EMOBON_[ObservatoryID]_[ARMS ID]_Ha_[dateIn]_[dateout] with dates given as YYMMDD
- **Location** Country, longitude, latitude, and depth range of each ARMS unit. Location accuracy is not defined, but for depth it should be better than a metre and for latitude, longitude it should be at least to two significant digits in the decimal values.
- **Dates** Dates that the units are emplaced and then later retrieved.
- **Description** A basic habitat description.

These metadata are added to the appropriate fields in the respective (site, sub-site, event, and material sample) pages in PlutoF. A guide to working with PlutoF for the ARMS project can be found in GitHub.

Image and Occurrence data are also uploaded to PlutoF.

- **Photographic images** of the ARMS unit plates as part of the sampling protocol, and also of any example specimens or field images. These can be in jpg or similar format. Of the images provided so far, a minimum of metadata are held in the image file (usually the image size, date, size), depending on the camera used to take the images. As these are so minimal, we do not harvest these metadata. As images come out of the embargo period, they are made publicly-viewable/downloadable from PlutoF directly.
- **Spreadsheets in which those images are described**: which plate they are of (1-9, counting from baseplate upwards, and Top or Bottom), whether these are of ARMS plate, specimens, or in the field. Template spreadsheets with fixed wording for each description are provided via the Handbook and in GitHub. Note that many

observatories did not upload these image description files, but many observatories did name the images following the image ID as described in the Handbook and this includes the plate information.

- **Spreadsheets of manual observations** of species: taken in the field, made while the ARMS units are disassembled, or observations of the photographs taken of the ARMS plates. These are also to be described following a fixed template provided in the Handbook and in GitHub.

The ARMS-MBON project in PluotF is kept behind a password-login, and is not yet made publicly-accessible. The images (image files, image spreadsheets) and manual observation spreadsheets, however, *are* made available for download directly from PlutoF after the embargo period.

PlutoF was used by ARMS-MBON while it was running under ASSEMBLE Plus and it continues to be used under EMO BON.

Genetic (meta)data. We are experimenting with using PlutoF to archive and share processed sequence data: ASVs produced by PEMA processing of the ARMS sequences, and species information arising from that processing. However, these (meta)data are not yet exported out of PlutoF and shared: when we do that, we will document the steps here.

3.2. The overview google sheet: observatory, event, sample, omics metadata

While running under ASSEMBLE Plus, ARMS-MBON used a google sheet in the ARMS google account to provide an overview of the events and material samples received to date, and their ENA accession numbers once the sequences have been uploaded there. The metadata in this google sheet include many of those also added to PlutoF, and to these are added some useful omics metadata.

- **IDs** Identifiers of the observatories, ARMS units, events, and material samples.
- **Location** Country, longitudes, latitudes, and contact details.
- **Dates** Dates that the units are emplaced and then later retrieved.
- **Description** Of the location of the ARMS units, including the type of monitoring area and habitat keywords, additionally information about which units are field replicates.
- **Material sample metadata** The sample fraction (motile or sessile), filter size (in microns), and the preservative used (after 2020, DMSO was fixed as the replicate and this metadatum is not longer mandatory). The material sample ID that was written on the falcon tube received by the sequencing centre (at HCMR or Paris) is also requested, as this often differed from the official MaterialSampleIDs.
- **Event metadata** Depth ranges of each ARMS unit, additional information on the field replicates and use of crate cover during retrieval.
- **Omics metadata** Being the ENA run accession numbers for the three gene types sequenced (18S, COI, ITS), for the samples themselves and their negative controls; some metadata on the sequence platform and QC tool used, initial number of paired-end reads and the FastQC report (pass, fail, warn). Note that the sequences themselves are always to be found in ENA, not in GitHub or on a google drive.

3.3. EMO BON logsheets: observatory, event, sample metadata

Under EMO BON, the observatories add their observatory, event, and sample metadata to their own ARMS “Hard-bottom sample” logsheets. These logsheets are formatted and standardised, and all observatories are required to conform to the instructions provided in the EMO BON Handbook and in the logsheets themselves. Mandatory fields include

- **IDs** Identifiers of the observatories, ARMS units, events, and material samples.
- **Location** Country, longitudes, latitudes, and institutional/personal contact details.
- **Dates** Dates that the units are emplaced and then later retrieved.
- **Descriptions** A set of ENVO terms to describe the locations and environment of the units.
- **Material sample metadata** The sample fraction (motile or sessile), filter size (in microns), and the preservative used. Metadata concerning the material sampling SOPs used, storage dates, durations, and temperature. Still to be added are specific information concerning any biobanking of material samples.
- **Event metadata** Depth ranges of each ARMS unit, additional information on the field replicates and use of crate cover during retrieval.

As no omics data have yet been fully processed by EMO BON, the management of omics metadata and data is yet to be formalised.

3.4. ENA

Our sequences are uploaded to ENA, from where they are available in its standard fastq format. The ARMS-MBON data gathered during the ASSEMBLE Plus period were given the following ENA accession codes:

- Each country is entered as a unique “Project” in ENA, receiving a unique **PRJ accession number**.
- Sequences from all sampling events are uploaded under each of those projects, receiving each **ERX (experiment) and ERR (run) accession numbers**.

Negative control sequences are also uploaded to ENA for each sequencing run and for each sampling event.

The run accession numbers for each of the three gene types (COI, 18S, ITS, and including the negative controls) are recorded in the ARMS overview google sheet. The MaterialSampleID that was written on the containers of the material samples are added to the ENA metadata, and as this not always exactly the same as the requested MaterialSampleID (following the ARMS-MBON Handbook, for example), the link between these two IDs can be made from their respective columns in the ARMS overview google sheet.

As there is no checklist that is applicable to ARMS units, we use the standard ENA checklist. This means that a minimum set of metadata are added: Instrument platform, model, and library layout, strategy, source, selection. The “sample title” field is filled with the material sample ID (see above). For each sample, the following 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 are tagged as "environmental_sample" and are submitted under the NCBI taxonomy ID 408172 corresponding to "marine metagenome".

The management of the omics (meta)data from ARMS-MBON during the EMO BON period will be documented in the EMO BON Handbook. The sequences will continue to be archived in ENA, but under a different set of accession numbers and with the EMO BON metadata.

3.4.1. Other omics (meta)data

Other metadata related to the processing to produce the sequences were kept by HCMR who processed the material samples from the ASSEMBLE Plus period of ARMS-MBON. and although these were not curated, they will be used when the data are published in biodiversity archives.

The material samples gathered during the EMO BON period of ARMS-MBON are processed by Genoscope following a set of procedures and metadata templates that were put together by the EMO BON secretariat. The curation of these (meta)data will be documented in the EMO BON Handbook and here, when those steps have been decided upon.

3.5. GitHub

Initially we used PlutoF and the overview google sheet as the sites for recording our ARMS-MBON data and metadata and for sharing within the project. Towards the end of the ASSEMBLE Plus period, it was decided that we needed to quality control, combine, semantically annotate, and share these data more widely, and we chose to do that via an ARMS-MBON GitHub space: https://github.com/arms-mbon for the GitHub-savvy, or this landing page otherwise.

In the (ASSEMBLE Plus) ARMS-MBON GitHub we have the following repositories

- **documentation**: for the Handbook, this DMP, SOPs, templates, and more
- **data_workspace**: for the data harvested from PlutoF and the overview google sheet
- **arms-mbon.github.io**: to hold the ARMS-MBON landing page contents
- **data_release_001** (etc): "frozen" datasets, taken from data_workspace, that hold data related to specific data publications; see their READMEs to learn more about these

All the work on harvesting, quality controlling, and combining the sampling, sequence, and species (meta)data is done inside the data_workspace repository. As data publications are created (e.g. associated with a scientific or data paper, a EurOBIS datasets, etc), these (meta)data are copied over into repositories called data_release_###.

Inside the data_workspace repository we manage the following data

- **QualityControlledData**: for the (meta)data downloaded from PlutoF and the metadata copied from the overview google sheet
    - *FromGS* are from the google sheet (all observatory, event, and sample metadata, including the ENA accession numbers)
    - *FromPlutoF* are from PlutoF (all observatory and event metadata and the download URLs for the image data)

    - *Combined* are the CSV files containing the combined PlutoF and google sheet metadata
    - *FromENA* is currently empty
- **ReformattedData**: for data destined for specific external sources, currently being the ARMS-MBON metadata in IMIS and the LifeWatch Tesseract workflow.
- **AnalysisData**: for any sequence analysis outputs, although this is currently empty.

Our GitHub repositories are formalised by being turned into Ro-Crate data packages: that is, by adding a formatted json file that describes the contents of the repositories (including of each folder therein), with appropriate semantics and provenance metadata, to make them machine-understandable.

ARMS-MBON running under EMO BON will also use GitHub as a place to QC, semantically annotate, and share the ARMS-MBON (and other) data. This GitHub space is still being developed and we will include the details here when they are available.

3.5.1. Harvest, quality control, combine

The harvest from PlutoF does the following

- All metadata and data in the ARMS-MBON account in PlutoF are harvested via a JSON dump into https://github.com/arms-mbon/data_workspace/tree/main/QualityControlledData/FromPlutoF.
    - This includes the **administrative, proceduring, sampling, image, and occurrence (meta)data**
    - Each Observatory is a separate folder, and inside each of these are five CSV files holding: an overview of all the events for that Observatory, the metadata from the Material Sample pages created for each Sampling Event, the metadata added to the Sequences and to the Observations pages, and the metadata for the Associated data (being images and spreadsheets) and including the download links for those data.
    - Note that we do not download the associated data, rather only their URLs: these data can be accessed by anyone from PlutoF using these URLs once we set their access conditions to “open”.
    - The same five spreadsheets containing the metadata for all observatories are also created.
    - Note that we do not do anything more with the metadata in the Sequences and Observations pages: these are only downloaded and not processed further.
- The download script applies a QC to the metadata, since it was found that the Observatories did not always use the correct IDs.
    - Observatory and ARMS IDs are converted to the correct names
    - Event IDs are constructed from the corrected Observatory and ARMS unit IDs and the provided event dates

The harvest from the google sheet does the following

- The observatory information and metadata tabs, the samples+sequences data and metadata tabs are copied into four CSV files in https://github.com/arms-mbon/data_workspace/tree/main/QualityControlledData/FromGS.

- A QC script is applied, checking for: consistency in the dates given with the dates in the IDs, correct identification of the observatory and ARMS unit IDs, consistency of the event and sample IDs with the individual pieces of metadata given in other columns.
- The QC script also compared the values in the google sheet with those from PlutoF, highlighting where there are differences.
- If the QC output indicates different, incorrect, or missing information, the data manager informs the observatories and requests that they update their information in the google sheet and PlutoF. Another round of harvesting from PlutoF and the googlesheet follows, until the QC is fully passed.

Finally, the metadata from PlutoF and the google sheet are combined into four CSV files and placed in https://github.com/arms-mbon/data_workspace/tree/main/QualityControlledData/Combined.

- Image metadata
- Observatory metadata
- Omics metadata
- Sampling event metadata

3.5.2. Semantic uplifting

The semantic uplifting of the four combined datasets – that is, adding machine-understandable descriptions taken from controlled vocabularies to each datum in the datasets and converting the CSV files into turtle format – is work that is still in progress. For the present we have the following semantic inputs

- The two metadata tabs from the google sheet, which contain the description, data type (taken from W3C's xsd), property (usually taken from schema.org), property URL (usually from the NERC vocabulary server), unit, and unit URL (also taken from NERC).
- A metadata CSV file that maps the various metadata names taken from PlutoF and the google sheet.

As we complete this work, it will be documented here.

3.7. IMIS

A metadata record for the ARMS-MBON data has been created in IMIS, and this will be updated until it contains all the data from ARMS-MBON that were gathered under ASSEMBLE Plus. This record links to data CSV files stored in the Marine Data Archive, and also links to the ARMS-MBON GitHub space.

4. Making our data FAIR

The FAIR principles are about making data **Findable**, **Accessible**, **Interoperable**, and **Re-usable**. These principles are explored in Wilkinson et al. (2016[2]), 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. Having said that, it is humans who will be exploring the data, and picking out what they want to work with, therefore it is important to provide F A I R for all possible human audiences. We have chosen to focus on the following audiences: the ARMS-MBON partners, other biotechnicians, and more general biodiversity scientists.

4.1. Making our 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 uses 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 exported in XML, JSON, and EML[3].

As stated above, a metadata record for the ARMS-MBON data has been created in IMIS, and this will be updated until it contains all the data from ARMS-MBON that were gathered under ASSEMBLE Plus. This record links to data CSV files stored in the Marine Data Archive, and also links to the ARMS-MBON GitHub space.

The following metadata fields are included in the ARMS record in IMIS:

- Title, abstract, and description which covers the content of the dataset
- Name and contact details for the person responsible for the dataset (including ORCIDs)
- A citation for the record
- Links to access the associated datasets
- Licence of use (BB CY)
- Keywords chosen for the dataset (taken from the ASFA vocabulary)
- The spatial and temporal 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 will also be created by IMIS via collaboration with DataCite once all the ASSEMBLE Plus ARMS data are in IMIS.

---

[2] 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*.

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

As the IMIS record includes a link to the GitHub space for ARMS-MBON, the additionally semantically-uplifted data are findable in this way, as well as directly via GitHub. We will also then publish the data in GBIF and/or EurOBIS in DarwinCore Archive format, but this step is still in progress.

For the ARMS-MBON data gathered under EMO BON, a similar metadata record will be created, but the details are still to be decided.

4.2. Making data Accessible

To make the ARMS data widely accessible,

- The data are provided as CSV files in the IMIS metadata record (download can be done via a click on the webpage or via the provided field in the EML, XML, or JSON version of the record) (human- and developer-accessible, machine accessible via IMIS's OAI-PMH protocol)
- The data will also be provided as turtle files (developer-accessible) in the Github space
- The sequences are accessible from ENA by using the ENA search/download methods and the accession numbers are also included in the above-mentioned CSV and turtle files
- The Data repository is also a Ro-Crate (developer- and machine-accessible)
- Once the data are published in GBIF and EurOBIS, they will be accessible via their search/download tools.
- Data are published only when they are open access, with a CC BY licence.

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

4.3. Making data Interoperable

Metadata are made interoperably by adhering to globally-accepted metadata standards. The IMIS datasets catalogue uses the EML metadata schema[4] (a standard for ecological data) as well as exporting in JSON and plain XML. GBIF also uses EML, 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 for taxonomy; and ASFA geoterms for geographic locations and for thematic keywords. The DarwinCore standard (for the data and metadata) are used by GBIF and EurOBIS.

The ARMS-MBON data in GitHub are provided in interoperable formats: CSV and turtle. Vocabularies used are: schema.org, W3C's xsd, those from BODC (from the NERC vocabulary server), ENVO, Marine Regions, ASFA, and a few others.

[4] 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.

4.4. Making data Re-usable

IMIS metadata records always provide open access to the metadata that are published, and the datasets there described. Any data (images, manual observations, sequences) that are still under the 1-year moratorium period will be closed access by being behind an account, but will always be CC BY thereafter. A citation for referencing the dataset is always included in the discovery metadata record. This moratorium period 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. Under EMO BON, this moratorium period is reduced to 6 months.

Data archived in the MDA remain there for as long as the project wishes and the MDA exists. The URLs pointing to the datasets therein are unique and persistent.

The data in GitHub will also remain in place, even if the project ends completely. Note that data related to a data/scientific publication/record will always be added to a data_release_### repository in our GitHub space and will be described there.

5. Ethical & GDPR aspects

The ARMS network has taken measures to be compliant with the EU regulations regarding the protection of personal data (http://ec.europa.eu/justice/data-protection/). Personal names and contact details are given in the datasets (e.g. observatory contacts) only upon approval.

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 (ASSEMBLE Plus until 2022, thereafter EMO BON). 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.

References -- not used but keep in case we do later

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.