EMBRC
EUROPEAN
MARINE
BIOLOGICAL
RESOURCE
CENTRE

European Marine Omics Biodiversity Observation Network (EMO BON)

Data Management Plan

EMBRC-ERIC, 4 Place Jussieu, Tour 46/00, 75005 Paris
www.embrc.eu

European Marine Omics Biodiversity Observation Network (EMO BON) – Data Management Plan

Ioulia Santi[1,2], Ibon Cancio[3], Erwan Corre[4], Cymon J. Cox[5], Katrina Exter[6], Mark Hoebeke[5], Anne Emmanuelle Kervella[1,4], Arnaud Laroquette[1], Christina Pavloudi[2], Marc Portier[6], Nicolas Pade[1]

1 European Marine Biological Resource Centre (EMBRC-ERIC) Headquarters, Paris, France

2 Hellenic Centre for Marine Research (HCMR), Institute of Marine Biology, Biotechnology and Aquaculture (IMBBC), Heraklion, Greece

3 Research Centre for Experimental Marine Biology and Biotechnology of Plentzia (PiE-UPV/EHU), University of the Basque Country (UPV/EHU), Plentzia, Spain

4 Sorbonne University, CNRS, Station Biologique de Roscoff, Roscoff France

5 Centro de Ciências do Mar (CCMAR), Universidade do Algarve, Faro, Portugal

6 Flanders Marine Institute (VLIZ), Ostende, Belgium

Version 1.0

February 2022

Contact:

Ioulia Santi

European Marine Biological Resource Centre
Paris Headquarters
4 Place Jussieu,
Tour 46/00, 1er étage, bureau 101
75252 Paris Cedex 05 (FR)
www.embrc.eu
Email: ioulia.santi@embrc.eu
Phone: +30 2 81 03 37 727
Mobile: +30 6 94 51 78 545

How to cite:

Ioulia Santi, Ibon Cancio, Erwan Corre, Cymon J. Cox, Katrina Exter, Mark Hoebeke, Anne Emmanuelle Kervella, Arnaud Laroquette, Christina Pavloudi, Marc Portier, Nicolas Pade. European Marine Omics Biodiversity Observation Network (EMO BON) – Data Management Plan. Version 1.0, April 2022. DOI: XXX

Table of contents

Introduction

In 2021 The European Marine Biological Resource Centre European Research Infrastructure Consortium (EMBRC-ERIC, hereinafter referred as EMBRC) launched the European Marine Omics Biodiversity Observation Network (EMO BON) which aims to be the first long-term marine genomic observatory on a European scale. EMO BON is a biological observatory network built on robust methodologies that produces quality-controlled genomic data following the high standards established by other networks such as ARGO[1] and Group on Earth Observations Biodiversity Observation Network[2] (GEO BON). EMO BON focuses on producing and providing access to long-term baseline genomic biodiversity data and supporting the monitoring of Essential Ocean and Biodiversity Variables (EOVs and EBVs) and ecosystem research.

EMO BON includes several locally-established genomic observatory projects at EMBRC partner institutions and connects them to one centrally-coordinated network. EMO BON adds value to the stations with support for long-term observation efforts and at the same time benefit from a wealth of standardised contextual data from the sampling sites. In addition, EMO BON encourages the establishment of new observatories at additional EMBRC stations, aiming for a broad geographic coverage of coastal Europe.

The governing body of EMO BON is the Operational Committee (OpCo) that is comprised of the EMBRC executive director, the EMO BON Data and Service Development Officer, the EMBRC Access and Benefit Sharing (ABS) Compliance Officer, the EMBRC International Cooperation Officer, one representative from each EMBRC node country, one representative from the EMBRC e-Infrastructure working group, and one representative from the EMBRC General Assembly. The EMBRC Access and Benefit Sharing (ABS) Compliance Officer and the EMBRC International Cooperation Officer act also as representatives of the EMBRC Access and Benefit-Sharing (ABS) working group. The OpCo oversees the function of EMO BON and decides on operations and the project’s development.

The EMO BON Handbook provides Standard Operational Procedures (SOPs) for sampling from the water column, soft substrate, and hard substrate habitats, thereby aiming to provide the baseline genomic data needed for a holistic view of the marine environment (Santi et al., 2021). Protocols have been developed or adopted from existing initiatives, and by consulting the expert scientists within EMBRC. The EMO BON Handbook is available from the Ocean Best Practices System (OBPS) repository[3] and it may also be found at the EMO BON webpage[4].

The project data, metadata, and their life cycle are described in this current document: the Data Management Plan (DMP). The DMP is a living document that will be updated continuously during a pilot period until the end of 2022 when the project is re-evaluated by the EMBRC General Assembly (GA) and the EMO BON OpCo. The **Genomic Data** generated by the project are released after a 6-month embargo period, starting with the data submission to a public database, set to allow the EMO BON operating stations to concretely plan on the data usage and to analyse and explore the data prior to public release. After this short time period, the data and metadata are released following

[1] ARGO: https://argo.ucsd.edu/
[2] GEO BON: https://geobon.org/
[3] EMO BON Handbook in OBPS: https://repository.oceanbestpractices.org/handle/11329/1738
[4] EMO BON webpage: https://www.embrc.eu/emo-bon

the FAIR data principles that the data are maximally Findable, Accessible, Interoperable and Reusable (Wilkinson et al., 2016).

EMO BON complies with all European regulations concerning the sampling of the marine ecosystems, the handling of samples and the data. This document is agreement with the EMBRC Rules of Operation approved by the EMBRC GA on March 2021 and revised (A6.2.3) by the EMBRC GA in November 2021. EMBRC infrastructure follows a DMP which is available on the EMBRC website[5]. This current DMP document comes to further specify the data policies and management procedures followed in EMO BON.

[5] EMBRC data management plan: https://www.embrc.eu/sites/default/files/publications/EMBRC_data_management_plan_V1_0.pdf

Data Description

Summary of EMO BON Datasets

EMO BON generates omics biodiversity data accompanied by rich metadata and complementary data that provide context and make the data relevant to different research fields. Specifically, EMO BON produces the following data:

- **Genomic Data**
  - **Metagenomic Raw Sequence Data (MetagRsd):** The original sequencing reads produced by the sequencing platform before any filtering or editing. Raw sequence data are in FASTQ file format (filename extension "fastq"), which is an ASCII text-based format for sequence read data and their corresponding sequence quality scores. This file format is the standard for storing the output of sequencing platforms. The project targets 50M reads for each sample, or approximately 14.9 Gb of data per sample.
  - **Metagenomic Quality-controlled Sequence Data (MetagQCsd):** The processed sequencing reads after the basic quality control of the metagenomic raw sequence data. Quality-controlled sequence data are stored in the FASTQ file format (filename extension "fastq"). The quality-controlled sequence data are less than 14.9 Gb per sample as they originate from the metagenomic raw sequence data. The final file size per sample depends on the quality control procedures.
  - **Metabarcoding Raw Sequence Data (MetabRsd):** The original metabarcoding reads as produced by the sequencing platform. Raw sequence data are in FASTQ file format (filename extension "fastq"). The project targets 1M reads for each sample and amplicon sequenced, that is approximately 298 Mb per sample.
  - **Metabarcoding Quality-controlled Sequence Data (MetabQCsd):** The processed sequencing reads after the basic quality control of the metabarcoding raw sequence data. Quality-controlled sequence data are in FASTQ file format (filename extension "fastq"). The quality-controlled sequence data are less than 298 Mb per sample as they originate from the metabarcoding raw sequence data. The final file size per sample depends on the quality control procedures. There is no universal concensus on the basic quality control of the metabacoding genetic data and therefore the quality control procedures performed are the minimum necessary (see section Genomic Data Quality Control).

- **Complementary (Meta)Data**

Complementary (Meta)data consist of environmental variables that are measured at the sampling site during a sampling event and are also considered part of the metadata that provide context for the genomic data. Many variables are included in the Essential Ocean Variables (EOVs) list maintained by The Global Ocean Observing System (GOOS)[6] and the Essential Biodiversity Variables (EBVs) produced by GEO BON[7]. The list of environmental variables measured by EMO BON is available in Appendix 1 (Table 5). Some variables are mandatory and are measured during every sampling event; five variables are mandatory for the water column (namely, sea surface

[6] EOVs: https://www.goosocean.org/index.php?option=com_content&view=article&id=14&Itemid=114

[7] EBVs: https://geobon.org/ebvs/what-are-ebvs/

temperature, subsurface temperature, chlorophyll, sea surface salinity and subsurface salinity) and four for the soft substrate (namely, sea surface temperature, temperature at the sediment's surface, pH, redox potential) samplings (see Appendix 1). Additionally, observatories may choose to measure any of the recommended variables. **Complementary (Meta)data** are collected at every sampling event, and they characterise and follow each sample collected at that sampling event. Upon measurement, **Complementary (Meta)data** are recorded in online e-log spreadsheets together with other metadata (see section Metadata) and afterwards they are converted to comma-separated delimited text files (CSV; filename extension "csv"). The size of the **Complementary (Meta)data** CSV files is not expected to be exceed 1 Mb for one habitat of one observatory.

- **Image Data** (only for the Hard Substrate habitat sampling events):
  - High resolution photographic pictures of settlement plates
  - High resolution photographic pictures of specimens for DNA metabarcoding
  - High resolution photographic pictures of habitat during sampling events

Image data are in Joint Photographic Experts Group format (filename extension "jpeg").

- **Occurrence Records** (only for the Hard Substrate habitat sampling events):
  - Abundance and coverage of species directly observed by morphological identification
  - Abundance and coverage of species observed from image data

Occurrence records are collected in CSV files (filename extension "csv").

For the management of the Hard Substrate Image Data, Occurrence Records, and metadata of EMO BON, the DMP of the Marine Biodiversity Observation Network for genetic monitoring of hard-bottom communities (Auronomous Reef Monitoring Structures Marine Biodiversity Observation Network, ARMS-MBON) is followed [8].

Sample Collection

Samples are collected at observatory stations widely distributed around the European coast (Figure 1). Three different coastal habitats are sampled within EMO BON: water column, soft substrate, and hard substrate. The individual observatories have selected which habitats to sample during the registration period in May 2021. The observatory stations of EMO BON and their chosen sampling habitats are listed in Table 1. Each sample is collected as four technical replicates to be used for different purposes; two of the technical replicates collected are processed and two are bio-banked. Data generated from the two processed replicates are used for cross-checking the results.

Sampling and basic sample processing takes place as described by the SOPs in the EMO BON Handbook (Santi et al., 2021), which provide all the necessary information for collecting and processing the water column, the soft substrate, and the hard substrate samples to the point of preparing them for DNA extraction (see also Appendix 2). The frequency of the sampling events is different for each habitat and biological community sampled and is summarised in Table 2. Samples are stored at the observatories until shipment for further analyses.

During sampling and sample processing, observatories collect metadata and record them in online e-log spreadsheets (see section Metadata). A unique identifier, the Source Material Identifier, is created for each collected sample according to the format described in the EMO BON Handbook.

[8] ARMS MBON DMP: https://www.assembleplus.eu/sites/assembleplus.eu/files/public/ARMS/ARMS_DMP.pdf

This identifier follows the sample through the various analysis steps, is attached to all the data produced and is used for linking the data and metadata. An additional traceability mechanism will be set to monitor each sample from collection to data production and publication. This mechanism will also include entries for the bio-banked samples. The traceability mechanism will be thoroughly described in a later version of the DMP.

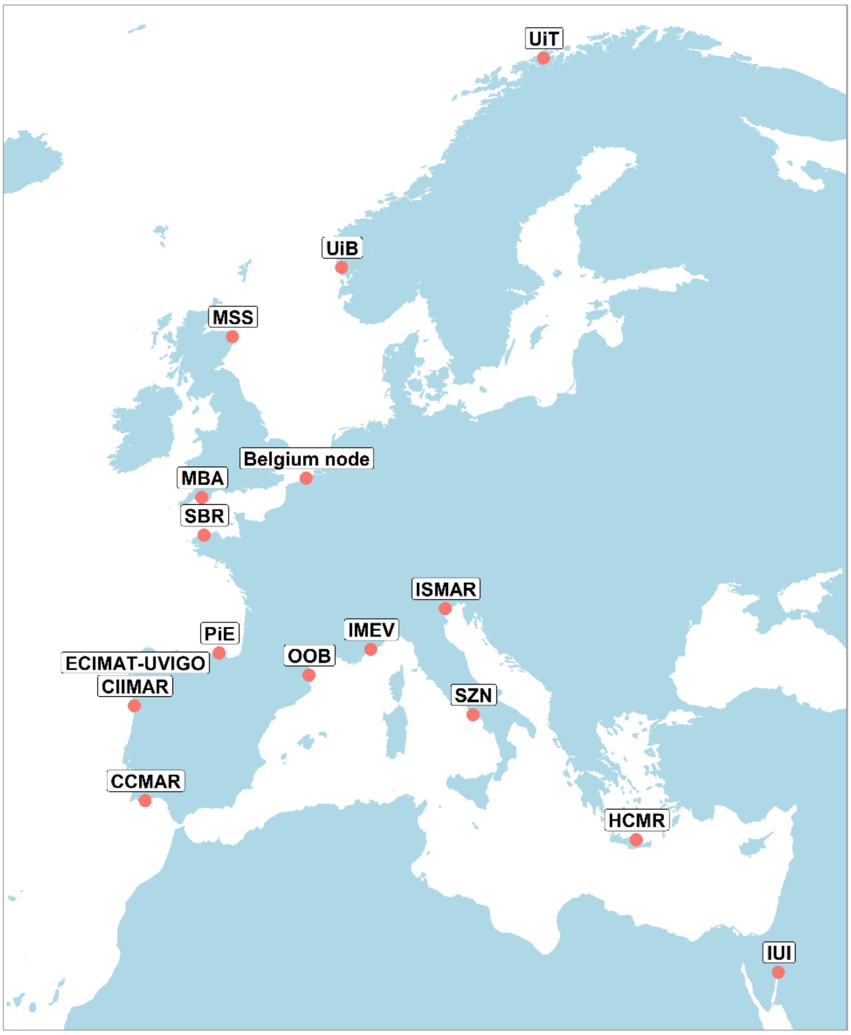

Figure 1: EMO BON observatory stations indicated on the European map, as of February 2022. The operator's name is shown for each observatory and abbreviations are explained in Table 1.

Table 1: List of EMO BON observatory stations and operating institutes. The habitats sampled by each station are summarized along with the SOPs followed. The SOPs are described in the EMO BON Handbook.

EMO BON station	Operator	Habitat(s)	SOP
EMOBON-BE	EMBRC Belgium (Belgium node): Ghent University, Flanders Marine Institute, Royal Belgian Institute of Natural Sciences and Katholieke Universiteit Leuven	Water Column Soft Substrates Hard Substrates	WaSOP 1 SoSOP 3 HaSOP
EMT CIM-UVIGO	Toralla Marine Science Station - Centre of Marine Research, University of Vigo (ECIMAT-UVIGO)	Water Column Soft Substrates Hard Substrates	WaSOP 1 SoSOP 3 (macrobenthos) HaSOP
PiE-UPV/EHU	Research Centre for Experimental Marine Biology and Biotechnology, Plentzia Marine Station (PiE)	Water Column Hard Substrates	WaSOP 1 HaSOP
Bay of Villefranche sur Mer southern France	Institut de la Mer de Villefranche (IMEV)	Water Column Hard Substrates	WaSOP 1 HaSOP
Roscoff Genomic Observatory	Station Biologique de Roscoff (SBR)	Water Column Soft Substrates Hard Substrates	WaSOP 1 SoSOP 3 HaSOP
HCMR-1-UBPC	Institute of Marine Biology, Biotechnology and Aquaculture - Hellenic Centre for Marine Research (HCMR)	Water Column Soft Substrates Hard Substrates	WaSOP 1 SoSOP 2 (microorganisms & macrobenthos) HaSOP
IUI	Interuniversity Institute for Marine Sciences in Eilat (IUI)	Water Column Hard Substrates	WaSOP 1 HaSOP
NEREAB	Stazione Zoologica Anton Dohrn (SZN)	Water Column Soft Substrates	WaSOP 1 SoSOP 3 (microorganisms)
Acqua Alta Oceanographic Tower	National Research Council, Institute of Marine Science (ISMAR)	Water Column Hard Substrates	WaSOP 1 HaSOP
Marineholmen	University of Bergen (UiB)	Water Column	WaSOP 1
Engenes Science Centre	The Arctic University of Norway (UiT)	Water Column	WaSOP 1
Ria Formosa	Centre for Marine Sciences (CCMAR)	Water Column Soft Substrates	WaSOP 1 SoSOP 1 (microorganisms)

Table 1 (continued).

EMO BON station	Operator	Habitat(s)	SOP
Porto	Interdisciplinary Centre of Marine and Environmental Research (CIIMAR)	Water Column	WaSOP 1
Plymouth Station L4	The Marine Biological Association (MBA)	Water Column	WaSOP 1
Stonehaven	Marine Scotland Science (MSS)	Water Column Soft Substrates	WaSOP 1 SoSOP 3 (macrobenthos)
Banyuls Observatory	Observatoire Océanologique de Banyuls (OOB)	Soft Substrates	SoSOP 2

Table 2: Sampling intervals for the different modules and sampled communities

Water Column	Once every two months (February, April, June, August, October, December)
Soft substrate - Microbial community	Once every two months (February, April, June, August, October, December)
Soft substrate - Meiobenthos community	Once every four months (April, June, October)
Soft substrate - Macrobenthos community	Twice a year (October, April)
Hard substrate (Autonomous Reef Monitoring Structures deployment / retrieval)	Flexible timeframe for operating stations

Data generation

Metagenomic samples from two consecutive sampling events are shipped to a centralised facility for DNA extraction and sequencing. Metabarcoding samples are shipped for DNA extraction and sequencing once a year. DNA extraction and sequencing follow the same procedures for all samples. The French National Sequencing Centre (Genoscope) undertakes the analyses of the samples (DNA extraction, purification, and sequencing).

Once the samples arrive at the sequencing centre, they are catalogued using an Analysis Reference Code. This reference code is a short one-word identifier that includes the project name and a 5-digit number. All reference codes are in the form “EMOBON00001”, and the number increases for each sample. This reference code is useful during the analysis workflow to reduce labelling mistakes during the laboratory work. Each of the analysis reference codes corresponds to one sample, hence to one Source Material Identifier. All Analysis Reference Codes, and the Source Material Identifier they correspond to are listed in spreadsheet files (Import Samples Spreadsheet) that are completed before each batch shipment. The Import Samples Spreadsheet files have been created by the sequencing centre and are completed by the EMBRC EMO BON Observation, Data and Service Development Officer prior to shipment (see section Administrative and logistic documents).

The **MetagRsd** and **MetabRsd** are generated by sequencing the samples and are uploaded to European Nucleotide Archive (ENA) by the sequencing centre (see section Data and Metadata Publication and ReleaseData and Metadata Publication and ReleaseData and Metadata Publication and Release). The **MetagRsd** are bioinformatically processed at the sequencing centre to perform the basic quality control; from this the **MetagQCsd** are produced. Similarly, after basic bioinformatic processing of the **MetabRsd**, the **MetabQCsd** are produced. The **MetagQCsd** and **MetabQCsd** are uploaded to ENA (see section Data and Metadata Publication and Release). With this flow of samples and data generation, there is a data upload once every four months for metagenomes and once a year for metabarcodes.

For the hard substrate samples, **Image Data** and **Occurrence Records** are transferred to the data management platform, PlutoF[9], where the necessary metadata are also added, as described in the ARMS-MBON DMP.

Genomic Data Quality Control

The first step for the **Genomic Data** quality control is performed by the Illumina NovaSeq sequencing platform itself, that is the Illumina Chastity Filter. Clusters of sequence reads are removed if they do not pass the filter. Only the sequence clusters that passed the Illumina Chastity Filter are included in the Raw Sequence Data fastq files and this is for both the metagenomic and the metabarcoding data). The sequence clusters that do not pass the Illumina Chastity Filter are discarded. This is a standard procedure performed by all Illumina platforms and is the default step to quality control performed during sequencing. All the Raw Sequence Data fastq files contain a quality score that is computed by the sequencing platform. There is a quality score for each base that is read during sequencing.

For the generation of the **MetagQCsd**, adapters, primers, and nucleotides of low sequence quality score are removed from both ends. Following this, sequences between the second unknown nucleotide and the end of the read are removed and sequence reads shorter than 30 nucleotides after trimming are discarded. Sequence reads that mapped onto run quality control sequences are removed. Finally, single reads are removed and kept in a separate fastq file. The clean sequencing reads that correspond to the ribosomal RNA genes are separated from other reads. After this bioinformatic quality control procedures the **MetagQCsd** are produced. The fastq filenames include the original filename of the raw sequence files and the suffix "_clean" (*_clean.fastq). This is a standard quality control procedure for metagenome sequences that allows the production of high quality, reliable data without contamination. The quality control is performed at the sequencing centre.

For the generation of the **MetabQCsd**, adapters and primers are removed from both ends. Following this, sequencing reads and their paired-reads that mapped onto run quality control sequences (PhiX genome) are removed. After this step, no further processing is performed as there is no consensus

[9] PlutoF Data Management and Publishing Platform: https://plutof.ut.ee/#/

on the steps that need to be performed. Therefore, future users can continue the analysis as desired. However, a comparison between the produced sequences and the negative controls (DNA extraction and PCR negative control) is computed and becomes accessible.

Documentation

Methodology for data collection

The collection of data follows the SOPs described the EMO BON Handbook (Santi et al., 2021). The EMO BON Handbook is a collective document that contains the methodologies for data collection. The EMO BON Handbook[10] is openly shared and discoverable through the OBPS repository under Creative Commons CC-BY licence[11]. The Handbook will be updated and supplemented with additional protocols depending on the development of the project. The methodologies for the collection of the Complementary (Meta)Data are recorded next to the Complementary (Meta)Data in online e-log spreadsheets together with other metadata (see section Metadata. In addition, the EMBRC guidelines for ABS were followed to ensure the compliance of using marine biological resources. The guidelines are published in a dedicated section of the EMBRC webpage[12].

Metadata

Extensive metadata are rigorously collected to accompany the sequence data generated in EMO BON. The metadata follow the Minimum Information for any x Sequence (MIxS) checklist v5.0[13] from the Genomic Standards Consortium (Yilmaz et al., 2011). Additional standards from the environmental packages *water* and *sediment* of the MIxS checklist have been used and the metadata terms have been cross-checked with the ENA MixS checklist[14] requirements. Extra fields have been added that capture details specific to the EMO BON samples and their provenance. The controlled vocabularies and ontologies used for specific metadata fields are described in Appendix 1 (Table 3 - Table 7)**Error! Reference source not found.**. Complete lists of the metadata may be found in Appendix 1 and in the online template spreadsheets: water column[15], soft substrates[16]. The hard substrate metadata terms are defined following the ARMS-MBON DMP. Metadata are separated in five categories to facilitate their collection and storage:

1. *Observatory Metadata*: Information regarding the observatory station, the sampling sites, their characteristics, and their representatives. *Observatory Metadata* are collected once for each observatory station, however, they characterise and follow all samples coming from one observatory.
2. *Sampling Metadata*: Detailed metadata recorded during each sampling event. All information on the when, who, and how a sample collection took place are included here. This category also includes information on the biobanking of samples. *Sampling Metadata* are collected at every sampling event and they characterise and label each collected sample during sequence production and the eventual release of the data.
3. *Complementary (Meta)data*: Environmental variables measured at the sampling site during a sampling event and the methodologies used to generate them. *Complementary (Meta)data* are also considered data as samples are collected and measured separately for their production.

[10] EMO BON Handbook in OBPS: https://repository.oceanbestpractices.org/handle/11329/1738
[11] Creative Commons licences: https://creativecommons.org/licenses/
[12] EMBRC ABS guidelines: https://www.embrc.eu/embrc-guides-abs
[13] MIxS chelist v5.0: https://gensc.org/mixs/
[14] ENA MIxS checklist: https://www.ebi.ac.uk/ena/browser/checklists
[15] EMO BON water column metadata online spreadsheet: https://docs.google.com/spreadsheets/d/1GWW5RN1veOZTQ1tynGtzuIpTAMyNmTygi15pvyKHucI/edit#gid=0
[16] EMO BON soft substrate metadata online spreadsheet: https://docs.google.com/spreadsheets/d/1iJsDXpP3AULxTnU5l7tJHPODR6U7IgDfjnxkndNUzcg/edit#gid=0

Complementary (Meta)data are collected at every sampling event and they characterise and label each collected sample to the production and the release of data.

An online e-log spreadsheet has been created for each observatory station and each sampled habitat for the recording of the *Observatory*, *Sampling* and *Complementary* (*Meta)data*. The metadata e-log sheets are organised so that each station completes the metadata on a separate spreadsheet. This way, erroneous edits and simultaneous working on the files is avoided. For example, *Observatory*, *Sampling* and *Complementary* (*Meta)data* for the water column samplings of observatory "NEREAB" are collected in one e-log spreadsheet and for the soft substrate samplings in a separate e-log spreadsheet. The online files are created in the EMBRC Headquarters Google account and are shared with the EMO BON contact persons of the observatory stations.

During the preparatory phase, the observatory stations completed registration forms with information on their stations and sampling sites. This information is used to prepare the *Observatory Metadata*. The observatories are to complete the *Sampling and Complementary (Meta)data* in the e-log spreadsheets after each sampling event. In communication with the observatories, the EMO BON Observation, Data and Service Development Officer is responsible for the curation of the metadata files. Following two sampling events, the *Observatory*, *Sampling* and *Complementary* (*Meta)data* e-log spreadsheets are downloaded and saved as CSV formatted files by the EMO BON Observation, Data and Service Development Officer.

The online e-log spreadsheets will later be converted to CSV files for publication.

4. *Analysis Metadata*: Description of the laboratory work and methodologies performed during DNA extraction and sequencing. These metadata are collected during each sample batch that is processed in the laboratory and they follow each sample to the production of the **Genomic Data** and the release of data.
5. *Post-sequencing Metadata*: Information regarding the processing of the raw sequence data (**MetagRsd** and **MetabRsd**) that result in the production of the quality-controlled sequence data (**MetagQCsd** and **MetabQCsd**). These metadata are collected during the computational processing of the raw data and are related only to the quality-controlled sequence sata (**MetagQCsd** and **MetabQCsd**).

The *Analysis* and the *Post-sequencing Metadata* are provided by the sequencing centre and are collected and organised in collaboration with the EMO BON Observation, Data and Service Development Officer. One spreadsheet is created for the documentation of the *Analysis* and *Post-sequencing Metadata* for each batch of samples analysed. The spreadsheets are converted to CSV files for publication.

The metadata term Source Material Identifier, which is the unique identifier for a collected sample, is included in the *Sampling, Analysis* and *Post-sequencing Metadata* and in the *Complementary (Meta)data* and is the link among all the produced metadata for a material sample.

Administrative and logistic documents

All documents created during the project related to EMO BON procedures and logistics are organised and backed up at frequent intervals (see section Storage and Backup). Documents related to EMO BON procedures and logistics include, but not limited to:

- Documents announcing the sampling events, the shipment of samples and other events.
- Instructions provided to the observatories for providing the metadata or for shipping the samples.

- Technical reports agreed between EMBRC and the sequencing centre. The technical reports include detailed information on what takes place after the samples are transferred to the sequencing centre (for example, storage temperature, methodologies, time schedules).
- Financial reports, quotations and invoices, issued for the duration of the project.
- Import Samples Spreadsheet that include the list of samples shipped for analysis. All analysis reference codes, and the Source Material Identifier they correspond to are listed in these spreadsheet files that are completed before each batch shipment. The file templates have been created by the sequencing centre and are completed by the EMO BON Observation, Data and Service Development Officer prior to shipment. This is a prerequisite for the shipment of samples.

EMO BON aims to comply with all national and international jurisdiction rules and regulations as detailed described in section Due Diligence, Legal Compliance and Ethics. Therefore, the following administrative documents that ensure the legal compliance of EMO BON are collected and stored (see section Storage and Backup): sampling permits and related documentation, documents related to complying with the Nagoya Protocol and the ABS requirements, legal agreements between EMBRC and the EMBRC node operators for each operating observatory station and Material Transfer Agreements (MTAs). Detailed description of the aforementioned documents can be found under section Due Diligence, Legal Compliance and Ethics.

Storage and Backup

Each batch of **Genomic Data** produced is initially stored by the sequencing centre and becomes available on a secure website for 1 month. Only the sequencing centre and EMBRC have access to the secure website. During this month, **Genomic Data** are uploaded to ENA by the sequencing centre. After ensuring the complete upload of the datasets, they are discarded from the sequencing centre storage spaces and from the secure website.

The **Image Data** and the **Occurrence Records** are stored and backed up locally by the individual operating stations. In addition, they are stored on the data management platform PlutoF and on are regular basis they are downloaded for long-term archiving in the Marine Data Archive[17] (MDA).

The ***Complementary (Meta)Data*** and the *Observatory* and *Sampling Metadata* files are located online in the EMBRC Google account and remain backed up in this space for 20 years after the end of the project. The online files are downloaded as CSV files and stored and backed up in the EMBRC Dropbox space. The *Analysis* and *Post-sequencing Metadata* files are also stored and backed up in the EMBRC Dropbox space. The metadata files are additionally archived in MDA for long-term storage.

Administrative documents, legal compliance documents and documents related to the various EMO BON procedures and logistics are stored in the EMBRC Dropbox folders and archived in the MDA.

[17] Marine Data Archive (MDA): http://www.vliz.be/en/marine-data-archive

Data and Metadata Publication and Release

All **Genomic Data** produced are submitted to ENA within a month after data generation by the sequencing centre. ENA Accession Numbers are given to each sequenced sample and to each sequencing run. The Sample Accession Number refers to all **Genomic Data** produced from one sample, and the Raw Reads Accession Number refers to the **Genomic Data** coming from one sequencing run. The ENA Accession Numbers and the Source Material Identifier are the identities of the material sample and the produced data. A unique Source Material Identifier may correspond to more than one ENA Accession Number(s) as more than one genomic dataset may came from one sample (see also section Appendix 2).

Genomic Data in ENA are organised hierarchically in *Studies*, *Samples* and *Runs*. One ENA *Study* includes one or more *Samples* that include one or more *Runs*. EMBRC has contacted the European Molecular Biology Laboratory European Bioinformatics Institute (EMBL-EBI) to discuss the creation of one more level to this organisation to hold “EMBRC EMO BON” as the umbrella project for the observatories. This way the EMO BON **Genomic Data** in ENA are organised hierarchically in *Project*, *Studies*, *Samples* and *Runs*. *Project* is “EMBRC EMO BON” and it includes several *Studies* that represent the observatories. One *Study* corresponds to one observatory, and it includes *Samples* (samples collected from this observatory) that subsequently include *Runs* (sequencing data that came from one sample). This improves the findability of EMO BON data and ensures that they are placed in one online space.

The metadata are submitted to ENA and to other chosen public portals (EuroOBIS, EMODnet Biology, and PANGEA) following two sampling events. The **Occurrence Records** are submitted to public databases (EuroOBIS, EMODnet Biology and PANGEA) together with the metadata. With the upload of the metadata to public databases, the same structure as with the ENA data is followed. The metadata are organised under the EMO BON umbrella project. Below this is the observatory level that includes all the metadata created by the observatory and the metadata records are organised by sampling event. This way, a Digital Object Identifier (DOI) is assigned to one sampling event from one observatory. The metadata are connected to the **Genomic Data** by the Source Material Identifier and ENA Accession Number(s). The ENA Accession Numbers are included in the *Post-sequencing Metadata*. Together with the metadata submission, a metadata record is created for the **Genomic Data** that clearly indicates the ENA Accession Numbers.

EMO BON intends to make all produced data open and fully available to the scientific community after the 6-month embargo period, that starts with the data submission to a public database. The **Image Data,** the **Occurrence Records,** and the metadata become open with their submission to the public databases. The release of the data and metadata is announced by frequent data publications that describe the datasets and their location. Through the data publications, the datasets acquire an additional DOI that marks the release of the data and metadata. The releases of the data and metadata are also announced in the EMO BON webpage.

Data and metadata are released under the most recent version of Creative Commons CC BY licence[18] and are available for redistribution with attribution to the original data publication.

[18] Creative Commons licences: https://creativecommons.org/licenses/

Due Diligence, Legal Compliance and Ethics

Genetic resources and data

EMO BON aims to comply with all national and international jurisdiction rules and regulations. EMBRC centralises, standardises, and coordinates due diligence and legal compliance procedures at the level of the infrastructure by taking actions regarding the two parallel legal frameworks that have been identified:

1. The activity of collecting marine samples falls under the United Nations Convention on the Law of the Sea (UNCLOS) that regulates the access to the Exclusive Economic Zone (EEZ) for marine scientific research performed within national jurisdiction. In this framework, an authorization (sampling permit) is necessary to access the maritime zone and sample for marine scientific research. Sampling permits are usually granted by the Maritime Affairs Department of each country to the marine stations. All observatory stations must ensure that they have sampling permits in place that allow them to perform sampling at the sites of interest. To assist the stations, EMBRC as the the coordinator of the network contacts the national authorities of each country to see what are the requirements attached to the sampling permits and if EMBRC could negotiate this centrally for all its stations. The sampling permits should cover EMO BON activities and the collected documentation should include details on the conditions attached to the permits. The related documentation is stored in the EMBRC Dropbox folders and archived in MDA (see also sections Administrative and logistic documents and Storage and Backup).
2. The way in which biological resources may be Accessed, and how the Benefits that result from their utilisation in research and development are Shared (ABS) with the country of origin refers to the ABS framework that follows the Nagoya Protocol. EMBRC, as the coordinator of the network contacts the national ABS focal points for the genetic resources under their respective sovereignty rights and declare that EMO BON intends to collect, store, and distribute samples genetic resources and produce digital sequence information (DSI) from those genetic resources. ABS related documentation is stored in the EMBRC Dropbox folders and archived in MDA (see also sections Administrative and logistic documents and Storage and Backup). ABS permits are complementary to sampling permits when applicable. More information on ABS are available in the EMBRC website[19].

In addition, the following actions related to the legal regulations attached to genetic resources, material samples and/or data of EMO BON will be taken during the project (see section Storage and Backup):

- Legal agreements (contracts) between EMBRC and the EMBRC node operators for each operating observatory station. This agreement will be the framework for all that take place within EMO BON and it will clarify the obligations and responsibilities of each party. The agreement will regulate all actions that will take place during the pilot phase and will be extendable if EMO BON continues after that period. This agreement will also include the MTA conditions for the transfer of material samples from the operating observatory stations to EMBRC Headquarters. They will outline all terms of use for the samples, their liability, what is permitted to be done with them, whether they can be re-distributed, and whether they should be destroyed after use.

[19] EMBRC ABS guidelines: https://embrc.eu/services/access-and-benefit-sharing

- MTA for the transfer of samples to entities outside the EMBRC. This standard MTA will outline the terms for using the samples outside the EMO BON network during their transfer to the sequencing center. The MTA will outline all terms of use for the samples, what is permitted to be done with them, whether they can be re-distributed, and whether they should be destroyed after use.
- The transfer of material samples will be managed by the traceability system that will be developed within EMO BON at the EMBRC level. This system will include information covering the full provenance of the resources and their life cycle to the production of data and digital sequence information. It will allow the direct tracing of any EMO BON material sample or digital data produced throughout the project. The implementing of such a traceability system will ensure the adherence to the FAIR data principles. The traceability mechanism will be described in a later version of the DMP after or during its development.

All Genomic Data and their metadata are owned by EMBRC being the entity responsible for their release. The **Complementary (Meta)Data**, **Image Data** and **Occurrence Records** collected are jointly owned by EMBRC and the operation stations. Data and metadata are released under the most recent version of Creative Commons CC BY licence and are available for redistribution with attribution to the original data publication. Data and metadata generated in EMO BON are accessible as described in section Data and Metadata Publication and Release.

Personal data

The personal data included in the metadata are names, email addresses and ORCIDs. The terms are listed here: *contact person name*, *contact person email, contact person ORCID, sampled by person (name), sampled by person ORCID, person responsible for storage (name), person responsible for storage ORCID, other person who contributed to the sample collection (name), other person who contributed to the sample collection ORCID*. Prior to the publication of the personal data with the metadata, the people involved will be asked to complete a General Data Protection Regulation (GDPR) form and indicate whether they give permission that their personal data will appear in public databases together with the metadata. This project does not involve human participants, therefore personal data will not be processed in any way. Personal data are included in the metadata as means of recognition to the people that contributed, future communication, and insurance that all procedures have been followed. In case metadata should appear online in any database, General Data Protection Regulation (GDPR) consent forms will be completed by the persons mentioned.

Data Management Responsibilities and Resources

EMBRC is responsible and liable for the overall observatory and its operation. EMBRC is responsible for the data management of EMO BON and for remotely monitoring all procedures.

The EMBRC Headquarter members that work on EMO BON are: EMBRC Executive Director, the Observation, Data and Service Development Officer, the ABS Compliance Officer and the International Cooperation Officer. EMBRC has calculated the time and financial resources of the EMO BON data management to be covered by the work of the Observation, Data and Service Development Officer. The OpCo of EMO BON, which is the governing body of the network, in collaboration with the EMBRC e-Infrastructure working group oversees the Data Management and may suggest the update of the Data Management Plan when necessary. The e-Infrastructure working group of EMBRC supports the implementation of the data management plan and provide guidance when necessary. The ABS working group of EMBRC supports the due diligence and the actions related to the legal compliance of EMO BON.

The tasks for the EMBRC Headquarters EMO BON team are:

- overseeing and curate the metadata collection in close collaboration with the operational stations
- the upload of the metadata to public databases in close collaboration with the EMBRC e-Infrastructure working group
- the effective connection of the data to their metadata by ENA Accession Numbers and by the Source Material Identifier
- the data publications that announce the release of the data
- contact the national focal points to negotiate the sampling permits and the ABS compliance for all the operating stations.
- ensure that resource and data management actions are compliant and procedures timely

Different entities are responsible for the data generation. The responsibilities will be further specified in the contracts to be drafted between EMBRC and EMBRC node operators for each observatory station. In brief, EMO BON operating stations are responsible for the sample collection and appropriate labelling. Operating stations are responsible for the samples and associated obligations until they are shipped to the EMBRC Headquarters. Once the samples have arrived at the EMBRC Headquarters and until their transfer to the French National Sequencing Centre (Genoscope), EMBRC is responsible for the samples and associated obligations. The sequencing centre is responsible for the sample laboratory analyses (DNA extraction and sequencing), the generation of the raw sequence data (**MetagRsd** and **MetabRsd**) and their upload to ENA. The sequencing centre is also responsible for the quality control of the raw sequence data, the production of the quality-controlled sequence data (**MetagQCsd** and **MetabQCsd**) and their upload to ENA.

EMBRC is using an Advanced Work Dropbox plan that is adequate to cover EMO BON needs for the pilot phase. There are no additional costs during the pilot phase.

The storage, archiving, publication, and release workflows described here undergo a pilot period coinciding to the EMO BON pilot phase. Therefore, the timelines and procedures are under testing and may take longer than stated in this current document.

FAIR data principles

Following “The FAIR Guiding Principles for scientific data management and stewardship” publication (Wilkinson et al., 2016), a large part of the scientific community is investing in making existing and newly produced data follow the four FAIR principles: Findable, Accessible, Interoperable, Reusable. Following FAIR data principles is not only a good practice that promotes research and gives more opportunities in the present but it is also a promise for the future. Findable, Accessible, Interoperable and Reusable data are the only way to access information on the past and present states of environment, predict changes, manage ecosystems and make sustainable decisions in the future.

The EMO BON **Genomic Data** follow the FAIR principles after the 6-month embargo period set upon them. The data and metadata become discoverable by both machines and humans. The procedures and conditions to access the data are clear and available to any user, including machines and humans. EMO BON data and metadata become interoperable for different applications, pipelines, programing languages and humans of various backgrounds. Finally, data and metadata can be used, replicated, transformed and applied to a different setting (The FAIR Cookbook). The following paragraphs summarise how EMO BON data and metadata follow FAIR principles.

Findable

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

A unique sample identifier (Source Material Identifier) is created for each sample collected according to the format described in the EMO BON Handbook. The identifier includes coded information on the observatory, the sampling date, the biological community sampled and the replicate number. Therefore, the identifier is unique for each collected sample. The Source Material Identifier is included in the metadata and characterises all the information related to a sample. During laboratory analyses, the identifier is translated to a shorter, only locally unique, Analysis Reference Code for simplicity during laboratory work, it is translated back to the unique identifier once the data are produced. Each Analysis Reference Code corresponds to one sample, hence to one Source Material Identifier. All Analysis Reference Codes, and the Source Material Identifier they correspond to are listed in spreadsheet files that are completed before each batch shipment for analyses and are archived to MDA. The metadata are connected to the **Genomic Data** by the Source Material Identifier and the ENA Accession Number(s) that are assigned to the data upon submission to ENA. The ENA Accession Number(s) are included in the *Post-sequencing Metadata*. Once metadata are submitted to PANGEA, a unique DOI is assigned to each submitted dataset. Similarly, metadata in EuroOBIS and EMODnet Biology get assigned a unique identifier.

EMO BON data are accompanied by rich and rigorous metadata that include, but are not limited to, information on the where, when, and how the samples were collected (*Observatory Metadata* and *Sampling Metadata*). Additional *Complementary (Meta)data* include the environmental variables measured during a sampling event and the methodologies used to collect measure them. Information on the laboratory analyses of the data, such as the DNA extraction method, the yields and the library preparation are collected as *Analysis Metadata*. The quality controlled bioinformatic procedures following sequencing to produce the Quality-controlled Sequence Data are documented as *Post-sequencing Metadata*. The Source Material Identifier is included in the metadata records and links together all the information collected as metadata. With the data submission to ENA, the ENA Accession Number(s) are added to the *Post-sequencing Metadata*.

The **Genomic Data** are submitted to ENA. The metadata and the **Occurrence Records** are submitted to searchable open public databases (EuroOBIS, EMODnet Biology, PANGEA). Part of the metadata are also submitted together with the data to ENA. A metadata record is created for each genomic dataset in the open public databases that clearly indicates where the Genomic Data are in ENA.

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

EMO BON data and metadata may be retrieved without any specialised protocol or procedure through the online databases. After the 6-month embargo, data are openly available in ENA and can be retrieved following the user-friendly procedures applied by ENA. Metadata are openly available in public databases and can be directly retrieved from there. ENA, as well as metadata public databases, provide easy access to data and good documentation to support the users. In addition, data are accessible directly to machines without manual human involvement. The releases of the data and metadata are also announced in the EMO BON webpage together with instructions on how to access them.

Metadata are stored in open public databases and are accessible for as long as the databases exist. The databases used are online, reliable and function in the long term, thus ensuring that metadata are accessible indefinitely.

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

Data and metadata are collected, stored, and published in readable forms that are easily converted into text files. **Genomic Data** are organised in fastq files, while **Occurrence Records** and metadata in csv files. The file formats of the data follow the rules and vocabularies of the fastq files. The **Image Data** for the hard substrates are in jpeg files respectively.

Controlled vocabularies and common ontologies are used for the metadata collection and explanations for each field are included in the metadata collection files that are available online. The metadata follow the MIxS checklist from the Genomic Standards Consortium (Yilmaz et al., 2011). The controlled vocabularies and ontologies used for specific metadata fields are described in Appendix 1 (Table 3 – 9).

Data and metadata are cross-referenced. Metadata fields refer to other metadata fields, aiming to provide as much clarity as possible in the information included.

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*

The metadata follow the MIxS checklist from the Genomic Standards Consortium (Yilmaz et al., 2011). Extra fields have been added that capture details specific to the EMO BON samples and describe their full provenance, including the sampling permits and the ABS required due diligence. **Genomic Data** are released after a 6-month embargo under Creative Commons CC BY licence and are available for usage to any user in the present and future. Metadata are released at 4-month time intervals and become immediately openly available under Creative Commons CC BY licence. Data publications announce the release of the dataset and the related metadata. The releases of the data and metadata are also announced in the EMO BON webpage. Full provenance information is provided by the metadata collection and the stored documentation.

References

Santi, I., Casotti, R., Comtet, T., Cunliffe, M., Koulouri, Y., Macheriotou, L., et al. (2021). European Marine Omics Biodiversity Observation Network (EMO BON) Handbook. Paris: Ocean Best Pactices System Repository doi:10.25607/OBP-1653.

The FAIR Cookbook FAIR Cookb. a Deliv. FAIRplus Proj. (grant Agreem. 802750), funded by IMI Program. a Priv. Partnersh. that Receiv. Support from Eur. Union's Horiz. 2020 Res. Innov. Program. EFPIA Co. Available at: https://fairplus.github.io/the-fair-cookbook/content/home.html.

Wilkinson, M. D., Dumontier, M., Aalbersberg, Ij. J., Appleton, G., Axton, M., Baak, A., et al. (2016). Comment: The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* 3, 1–9. doi:10.1038/sdata.2016.18.

Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J. R., Amaral-Zettler, L., et al. (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat. Biotechnol.* 29, 415–420. doi:10.1038/nbt.1823.

Appendix 1: List of Metadata

EMO BON metadata, relevant standards and vocabularies are listed in the following tables. Full descriptions of the terms and the structure of the metadata spreadsheet files may be found online: water column[20], soft substrates[21].

Table 3: List of EMO BON Metadata (Category: Observatory). EMO BON terms are metadata terms that were introduced in this project and are not relevant to any know metadata standards and vocabularies. These terms are useful for different procedures within this network such as the traceability of samples or the recognition of the people and institutions that participate. EMO BON terms were defined during the registration of the observatories or in the EMO BON Handbook or are described in the metadata spreadsheet files.

Metadata term	***Relevant standards and vocabularies***
project name	MIxS checklist v5.0[22]
geographic location (latitude)	MIxS checklist v5.0
geographic location (longitude)	MIxS checklist v5.0
geographic location (country)	MIxS checklist v5.0 & INSDC country list [23]
observatory location ocean or sea	Marine Regions[24]
observatory regional location	Marine Regions
observatory local location	Marine Regions
environment (biome)	MIxS checklist v5.0 & Environment Ontology (ENVO)[25]
environment (feature)	MIxS checklist v5.0 & Environment Ontology (ENVO)
environmental package	MIxS checklist v5.0
total depth of water column	MIxS checklist v5.0
site distance from shore	EMO BON term
organization of the observatory station	EMO BON term
country of the organization of the observatory station	INSDC country list
EDMO id of the operating organization	European Directory of Marine Organizations (EDMO)[26]
observatory id	EMO BON term
water column sampling site id [a]	EMO BON term
soft substrate sampling site id [b]	EMO BON term
extra sampling site information	EMO BON term
contact person name	EMO BON term
contact person email	EMO BON term
contact person orcid	EMO BON term
sediment type [b]	MIxS checklist v5.0

[a] Only in the water column habitat metadata; [b] Only in the soft substrate habitat metadata

[20] EMO BON water column metadata online spreadsheet: https://docs.google.com/spreadsheets/d/1GWW5RN1veOZTQ1tynGtzuIpTAMyNmTygi15pvyKHucI/edit#gid=0

[21] EMO BON soft substrate metadata online spreadsheet: https://docs.google.com/spreadsheets/d/1iJsDXpP3AULxTnU5l7tJHPODR6U7IgDfjnxkndNUzcg/edit#gid=0

[22] MIxS chelist v5.0: https://gensc.org/mixs/

[23] INSDC country list: http://insdc.org/country.html

[24] Marine Regions: https://www.marineregions.org/

[25] Environment Ontology (ENVO): http://www.ontobee.org/ontology/ENVO

[26] European Directory of Marine Organizations (EDMO):https://edmo.seadatanet.org/search

Table 4: List of EMO BON Metadata (Category: Sampling). EMO BON terms are metadata terms that were included in this project and are not relevant to any know metadata standards and vocabularies. These terms are useful for different procedures within this network such as the traceability of samples or the recognition of the people and institutions that participate. EMO BON terms were defined during the registration of the observatories or in the EMO BON Handbook or are described in the metadata spreadsheet files.

Metadata term	***Relevant standards and vocabularies***
source material identifiers	MIxS checklist v5.0 & EMO BON term
investigation type	MIxS checklist v5.0
environment (material)	MIxS checklist v5.0 & Environment Ontology (ENVO)
collection date	MIxS checklist v5.0
sampling event	EMO BON term
sampled by person (name)	EMO BON term
sampled by person orcid	EMO BON term
tidal stage	MIxS checklist v5.0
geographic location (depth)	MIxS checklist v5.0
amount or size of sample collected	MIxS checklist v5.0
time to filtration [a]	EMO BON term
size fraction selected [a]	MIxS checklist v5.0
size-fraction lower threshold	MIxS checklist v5.0
size-fraction upper threshold	MIxS checklist v5.0
membrane filter cut [a]	EMO BON term
sample collection device or method	MIxS checklist v5.0
sample material processing	MIxS checklist v5.0
sample material processing deviations	EMO BON term
sample storage duration	MIxS checklist v5.0
sample deposit date	EMO BON term
sample storage location	MIxS checklist v5.0
sample storage temperature	MIxS checklist v5.0
other person who contributed to the sample collection (name)	EMO BON term
other person who contributed to the sample collection orcid	EMO BON term
person responsible for storage (name)	EMO BON term
person responsible for storage orcid	EMO BON term
sample shipment date	EMO BON term

[a] Only in the water column habitat metadata

Table 5: List of EMO BON Complementary (Meta)Data. EMO BON terms are metadata terms that were included in this project and are not relevant to any know metadata standards and vocabularies. These terms are useful for different procedures within this network such as the traceability of samples or the recognition of the people and institutions that participate. EMO BON terms were defined during the registration of the observatories or in the EMO BON Handbook or are described in the metadata spreadsheet files.

(Meta)Data term	*Relevant standards and vocabularies*
chlorophyll *	MIxS checklist v5.0 & EOV
chlorophyll method	EMO BON term & Darwin Core term[27]
sea surface temperature * [¶]	MIxS checklist v5.0 & EOV
sea surface temperature method	EMO BON term & Darwin Core term
sea subsurface temperature *	MIxS checklist v5.0 & EOV
sea subsurface temperature method	EMO BON term & Darwin Core term
sea surface salinity *	MIxS checklist v5.0 & EOV
sea surface salinity method	EMO BON term & Darwin Core term
sea subsurface salinity *	MIxS checklist v5.0 & EOV
sea subsurface salinity method	EMO BON term & Darwin Core term
alkalinity	MIxS checklist v5.0
alkalinity method	EMO BON term & Darwin Core term
ammonium	MIxS checklist v5.0
ammonium method	EMO BON term & Darwin Core term
atmospheric data	MIxS checklist v5.0
bacterial carbon production	MIxS checklist v5.0
bacterial carbon production method	EMO BON term & Darwin Core term
bacterial production [a]	MIxS checklist v5.0
bacterial production method [a]	EMO BON term & Darwin Core term
bacterial respiration	MIxS checklist v5.0
bacterial respiration method	EMO BON term & Darwin Core term
biomass	MIxS checklist v5.0
biomass method	EMO BON term & Darwin Core term
chemical administration	MIxS checklist v5.0
conductivity [a]	MIxS checklist v5.0
conductivity method [a]	EMO BON term & Darwin Core term
density [a]	MIxS checklist v5.0
density method [a]	EMO BON term & Darwin Core term
dissolved carbon dioxide	MIxS checklist v5.0
dissolved carbon dioxide method	EMO BON term & Darwin Core term
dissolved hydrogen	MIxS checklist v5.0
dissolved hydrogen method	EMO BON term & Darwin Core term
dissolved inorganic carbon	MIxS checklist v5.0
dissolved inorganic carbon method	EMO BON term & Darwin Core term

[27] Darwin Core terms: https://dwc.tdwg.org/list/#dwc_measurementMethod

dissolved organic carbon	MIxS checklist v5.0
dissolved organic carbon method	EMO BON term & Darwin Core term
dissolved organic nitrogen	MIxS checklist v5.0
dissolved organic nitrogen method	EMO BON term & Darwin Core term
downward PAR [a]	MIxS checklist v5.0
downward PAR method [a]	EMO BON term & Darwin Core term
light intensity [a]	MIxS checklist v5.0
light intensity method [a]	EMO BON term & Darwin Core term
dissolved oxygen	MIxS checklist v5.0
dissolved oxygen method	EMO BON term & Darwin Core term
magnesium [b]	MIxS checklist v5.0
magnesium method [b]	EMO BON term & Darwin Core term
methane [b]	MIxS checklist v5.0
methane method [b]	EMO BON term & Darwin Core term
n-alkanes	MIxS checklist v5.0
n-alkanes method	EMO BON term & Darwin Core term
nitrate	MIxS checklist v5.0
nitrate method	EMO BON term & Darwin Core term
nitrite	MIxS checklist v5.0
nitrite method	EMO BON term & Darwin Core term
organic carbon	MIxS checklist v5.0
organic carbon method	EMO BON term & Darwin Core term
organic matter	MIxS checklist v5.0
organic matter method	EMO BON term & Darwin Core term
organic nitrogen	MIxS checklist v5.0
organic nitrogen method	EMO BON term & Darwin Core term
organism count	MIxS checklist v5.0
organism count method	EMO BON term & Darwin Core term
oxygenation status of sample [b]	MIxS checklist v5.0
pH [¶]	MIxS checklist v5.0
pH method	EMO BON term & Darwin Core term
particulate organic carbon	MIxS checklist v5.0
particulate organic carbon method	EMO BON term & Darwin Core term
particulate organic nitrogen	MIxS checklist v5.0
particulate organic nitrogen method	EMO BON term & Darwin Core term
particle classification [b]	MIxS checklist v5.0
perturbation	MIxS checklist v5.0
petroleum hydrocarbon	MIxS checklist v5.0
petroleum hydrocarbon method	EMO BON term & Darwin Core term
phaeopigments	MIxS checklist v5.0

phaeopigments method	EMO BON term & Darwin Core term
phosphate	MIxS checklist v5.0
phosphate method	EMO BON term & Darwin Core term
pigments	MIxS checklist v5.0
pigments method	EMO BON term & Darwin Core term
porosity [b]	MIxS checklist v5.0
pressure	MIxS checklist v5.0
pressure method	EMO BON term & Darwin Core term
primary production	MIxS checklist v5.0
primary production method	EMO BON term & Darwin Core term
redox potential [¶]	MIxS checklist v5.0
redox potential method	EMO BON term & Darwin Core term
sediment temperature at the top 0-5 cm [¶ b]	EMO BON term
sediment temperature at the top 0-5 cm method [b]	EMO BON term & Darwin Core term
silicate	MIxS checklist v5.0
silicate method	EMO BON term & Darwin Core term
sulfate	MIxS checklist v5.0
sulfate method	EMO BON term & Darwin Core term
sulfide	MIxS checklist v5.0
sulfide method	EMO BON term & Darwin Core term
total carbon	MIxS checklist v5.0
total carbon method	EMO BON term & Darwin Core term
total dissolved nitrogen	MIxS checklist v5.0
total inorganic nitrogen	MIxS checklist v5.0
total nitrogen concentration	MIxS checklist v5.0
total nitrogen concentration method	EMO BON term & Darwin Core term
total particulate carbon	MIxS checklist v5.0
total particulate carbon method	EMO BON term & Darwin Core term
total phosphorus	MIxS checklist v5.0
total phosphorus method	EMO BON term & Darwin Core term
turbidity [a]	MIxS checklist v5.0
water content [b]	MIxS checklist v5.0
water current	MIxS checklist v5.0
water current method	EMO BON term & Darwin Core term
miscellaneous parameter	MIxS checklist v5.0
miscellaneous parameter method	EMO BON term & Darwin Core term

* Mandatory variable for water column habitat; [¶] Mandatory variable for soft substrate habitat; [a] Only in the water column habitat metadata; [b] Only in the soft substrate habitat metadata

Table 6: List of EMO BON Metadata (Category: Analysis). EMO BON terms are metadata terms that were included in this project and are not relevant to any know metadata standards and vocabularies. These terms are useful for different procedures within this network such as the traceability of samples or the recognition of the people and institutions that participate. EMO BON terms were defined during the registration of the observatories or in the EMO BON Handbook or are described in the metadata spreadsheet files.

source material identifiers analysis	EMO BON term
nucleic acid extraction	MIxS checklist v5.0
nucleic acid amplification	MIxS checklist v5.0
library size	MIxS checklist v5.0
library reads sequenced	MIxS checklist v5.0
library layout	MIxS checklist v5.0
library vector	MIxS checklist v5.0
library screening strategy	MIxS checklist v5.0
target gene	MIxS checklist v5.0
target subfragment	MIxS checklist v5.0
pcr primers	MIxS checklist v5.0
multiplex identifiers	MIxS checklist v5.0
adapters	MIxS checklist v5.0
pcr conditions	MIxS checklist v5.0
sequencing method	MIxS checklist v5.0
whole genome amplification approach	MIxS checklist v5.0
whole genome amplification kit	MIxS checklist v5.0
sample volume or weight for DNA extraction	MIxS checklist v5.0
dna concentration after extraction	EMO BON term
dna concentration after extraction method	EMO BON term
sample arrival date	EMO BON term
reference code sequencing centre	EMO BON term
analysis	EMO BON term

Table 7: List of EMO BON Metadata (Category: Post-sequencing). EMO BON terms are metadata terms that were included in this project and are not relevant to any know metadata standards and vocabularies. These terms are useful for different procedures within this network such as the traceability of samples or the recognition of the people and institutions that participate. EMO BON terms were defined during the registration of the observatories or in the EMO BON Handbook or are described in the metadata spreadsheet files.

sequence quality check	MIxS checklist v5.0
chimera check	MIxS checklist v5.0
16S recovered	MIxS checklist v5.0
16S recovery software	MIxS checklist v5.0
number of standard tRNAs extracted	MIxS checklist v5.0
tRNA extraction software	MIxS checklist v5.0
completeness approach	MIxS checklist v5.0
contamination score	MIxS checklist v5.0
contamination screening input	MIxS checklist v5.0
contamination screening parameters	MIxS checklist v5.0
decontamination software	MIxS checklist v5.0
binning parameters	MIxS checklist v5.0
binning software	MIxS checklist v5.0
reassembly post binning	MIxS checklist v5.0
MAG (Metagenome-Assembled Genomes) coverage software	MIxS checklist v5.0
taxonomic identity marker	MIxS checklist v5.0
assembly name	MIxS checklist v5.0
assembly quality	MIxS checklist v5.0
assembly software	MIxS checklist v5.0
annotation	MIxS checklist v5.0
relevant electronic resources for the sequencing work	EMO BON term
relevant electronic resources for the post-sequencing work	EMO BON term
relevant standard operating procedures for the sequencing work	EMO BON term
relevant standard operating procedures for the post-sequencing work	EMO BON term
number of contigs	MIxS checklist v5.0
number of reads	MIxS checklist v5.0
reference database(s)	MIxS checklist v5.0
similarity search method	MIxS checklist v5.0
taxonomic classification	MIxS checklist v5.0
completeness score	MIxS checklist v5.0
completeness software	MIxS checklist v5.0
sample accession number	MIxS checklist v5.0
raw reads accession number	MIxS checklist v5.0

Appendix 2: Samples Summary

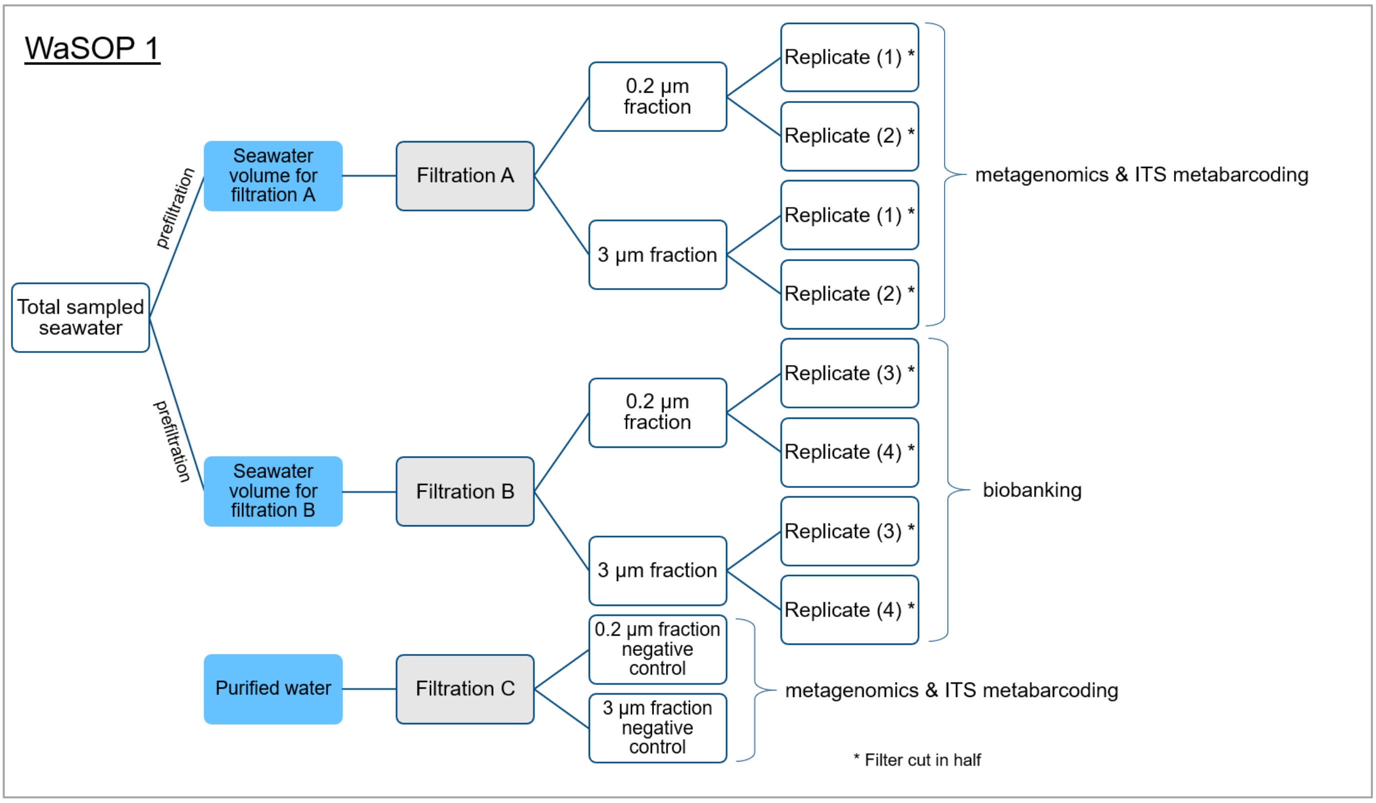

Figure 2: Diagram summarizing the WaSOP1 procedures, the samples collected, and the destination of the samples after collection, according to the EMO BON Handbook.

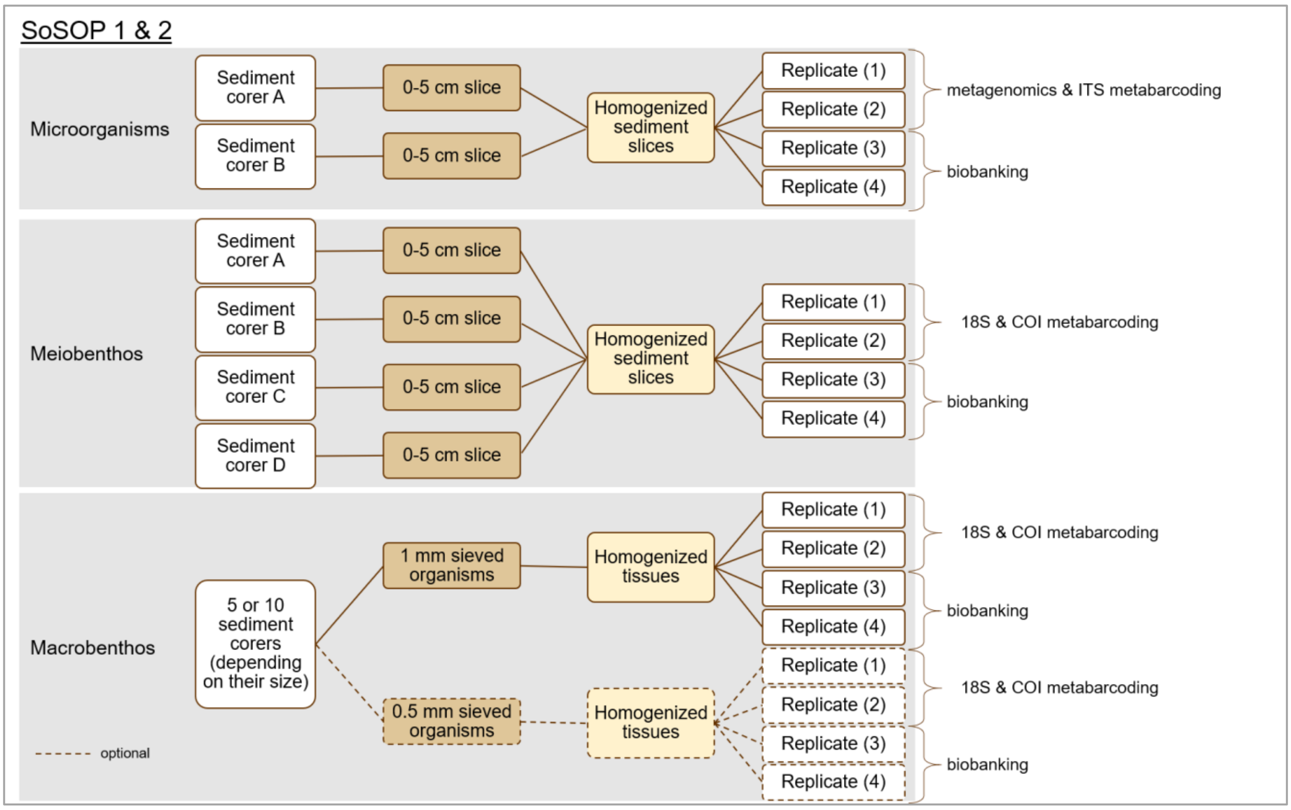

Figure 3: Diagram summarizing the SoSOP1 and SoSOP2 procedures, the samples collected, and the destination of the samples after collection, according to the EMO BON Handbook. The dashed lines and boxes in the macrobenthos collection diagram indicate the sequential sieving performed optionally by the stations that usually use smaller sieve size for macrobenthos collection.

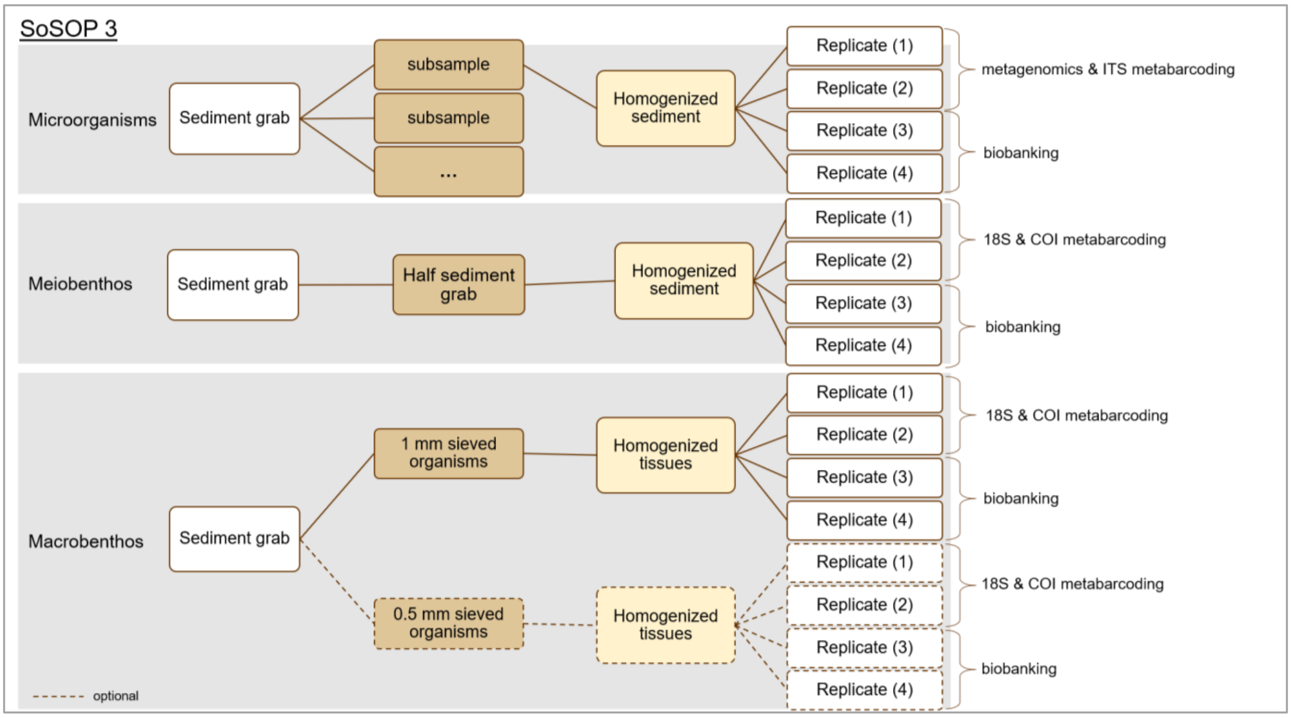

Figure 4: Diagram summarizing the SoSOP3 procedures, the samples collected, and the destination of the samples after collection, according to the EMO BON Handbook. The dashed lines and boxes in the macrobenthos collection diagram indicate the sequential sieving performed optionally by the stations that usually use smaller sieve size for macrobenthos collection.