ASSEMBLE
ASSOCIATION OF EUROPEAN MARINE BIOLOGICAL LABORATORIES EXPANDED

Ocean Sampling Day
Data Management Plan

Authors: Katrina Exter (VLIZ), Dimitra Mavraki (HCMR), Melathia Stavroulaki (HRMC), Jon Bent Kristoffersen (HCMR), Georgios Kotoulas (HCMR)
Date: Oct 31 2022

Scope of this document

Ocean Sampling Day (OSD) is one of the Genomics Observatories managed by Joint Research Activity of ASSEMBLE Plus. The Data Management Plan (DMP) of ASSEMBLE Plus is the inspiration for this OSD DMP. The management of the data arising from the water samples gathered by the OSD participants – the logsheet and eDNA data – to final publication in the IMIS datasets catalogue (Integrated Marine Information System), is described here. Subsequent publication steps which are still in progress – and which will only be completed after the ASSEMBLE Plus project has ended – will also be summarised; to be updated at the appropriate point.

The OSD data covered by this document are those from OSD2014 (which preceded the ASSEMBLE Plus project), and OSD2018 and OSD2019 (which were carried out under ASSEMBLE Plus).

- OSD 2014 was managed by the Marine Microbial Biodiversity, Bioinformatics, Biotechnology (MicroB3) project (https://www.microb3.eu/osd.html), but as we have added these data to our ASSEMBLE Plus OSD outputs, a description of how those data were managed and made FAIR by MicroB3 is included here.
- ODS2018 and OSD2019 data are completed and published, and the data life-cycle and FAIRification are explained here.
- OSD2020 and 2021 were interrupted by Covid, and while material samples were collected and shipped, they have not all yet been sequenced and hence not published. When those OSD datasets are produced, they will be managed in the same way as the OSD2018 and 2019 data.

In this DMP we explain how we have made the OSD data FAIR

- **Findable**: data can be found online by searching in data catalogues or portals. Keywords are descriptive.
- **Accessible**: data can be accessed by humans and programmatically from those data catalogues or portals.
- **Interoperable**: data are provided in interoperable file formats and with a structure and organisation that is developer-accessible as well as human-readable. Metadata that are taken from controlled vocabularies are used to annotate the data and their metadata records (in those catalogues or portals where they are made Findable).
- **Re-usable**: data have a clear licence and sufficient provenance that they could be reproduced.

Summary of the OSD material and data life-cycle

The order of activities for each OSD year from 2018 onwards are summarised here.

1. A new OSD sampling event (occurring always around June 21st) is announced on the OSD webpage and the Handbook, protocols, and permits templates for that year are there provided. These documents varied only very slightly from year to year.
2. The sampling events happened. Samples were collected, processed, and shipped to HCMR (Hellenic Centre for Marine Research) following the protocols documented in the Handbook and in the two SOPs (standard operating procedures). Logsheets were shipped with the samples or sent via email.
3. The logsheets were manually copied into a google sheet by HCMR, and this included some manual QC steps.
4. Once all the samples were received at HCMR, they were sequenced (16S, 18S, and shotgun metagenomics; at HCMR and by Genoscope)
5. The sequences were uploaded to ENA (the European Nucleotide Archive), receiving project, study, experiment, and run accession numbers.
6. The run accession numbers were added to the google sheet created from the logsheets data.
7. The google sheet was downloaded as CSV and placed on the OSD GitHub repository. There the data underwent a final QC, were semantically annotated, and provided for open access in CSV and turtle format.
8. The CSV files and documentation were uploaded to the Marine Data Archive and from there linked to their respective IMIS metadata records (2018, 2019).

Data access points:

- OSD2018 metadata record in the IMIS datasets catalogue, from where the documentation for 2018 and the CSV file with the OSD data can be downloaded
- OSD2019 metadata record in the IMIS datasets catalogue, from where the documentation for 2019 and the CSV file with the OSD data can be downloaded
- OSD2014 metadata record in the IMIS datasets catalogue. This is a copy of the PANGAEA metadata record and provided only so that data are in the same format as the OSD data collected under ASSEMBLE Plus

- OSD GitHub repository, providing access to the OSD2018, 2019, 2014 data, and to the documentation and SOPs.

Details for the OSD data

Logsheet (meta)data

The (meta)data collected by the sampling scientists of any one year of OSD were processed in the following way:

- Logsheets from the sampling stations. Paper logsheets were received from each sampling station and these were scanned and these digital copies stored in the MDA for the ASSEMBLE Plus years, and in PANGAEA for OSD2014. All logsheet scans have a CC BY licence and can be accessed via their respective metadata records (see below). These logsheets followed the template provided in the Handbook of OSD, which in turn is based on the handbook written by MicroB3 for OSD 2014: hence, the event, sample, and environmental values requested were the same for OSD2014 and OSD2018/9. This makes the original logsheets **Findable, Accessible**, and **Re-usable** (being PDFs, they are not considered to be very Interoperable).
- Logsheets combined. The logsheets for the ASSEMBLE Plus OSD years were manually copied into google sheets, one per OSD year. Use of google sheets was to allow for easy co-working between HCMR (who were running OSD) and VLIZ (who were doing the data management). However, these working spreadsheets are not intended to be used by others and therefore these data are closed access.
- Logsheets combined and standardised.
  - OSD2018 and 2019: The google sheets were copied over into CSV files and placed in the OSD GitHub repository. For each data CSV file, we provide a metadata CSV file where the column titles are semantically annotated with the data type, property and property URL (e.g. to the relevant BODC vocabularies for the environmental parameters and schema.org for event metadata), and with the unit property and property URL. The data are also provided in turtle format for developer-friendly access. This makes the OSD2018 and OSD2019 data **Interoperable**.
  - OSD2014: We downloaded the TSV file containing the OSD2014 data from the PANGAEA record, and then also provide these data in the OSD GitHub repository with a metadata CSV file (with semantic annotations) and in turtle format. This makes the OSD2014 data better **Interoperable** with the OSD2018 and OSD2019 data.

  All data are open access (MIT licence on GitHub and CC BY on the metadata records).
- Documentation and protocols. The Handbook, sample and molecular SOPs, the template Material Transfer Agreement (MTA) to be signed, the data policy to be signed, and the Access and Benefit Sharing agreement (to be signed). They can be found via the OSD github account. All documentation is open access.

A summary of the metadata that were requested on the logsheets, and their semantic values, are listed below.

ColumnTitle	Datatype	Property	ObservablePropertyUrl	Unit	UnitUrl
EventID	xsd:string	schema:identifier			
SampleID	xsd:string	schema:identifier			
SampleID_ENA	xsd:string	schema:identifier			
SiteID	xsd:string	schema:identifier			
DateTime	xsd:datetime	schema:startDate			
Latitude	xsd:double	schema:latitude		decimal degree	http://vocab.nerc.ac.uk/collection/P06/current/UAAA/
Longitude	xsd:double	schema:longitude		decimal degree	http://vocab.nerc.ac.uk/collection/P06/current/UAAA/
environment (biome)	xsd:string	schema:location			
environment (material)	xsd:string	schema:description			
environment (feature)	xsd:string	schema:description			
project name	xsd:string	schema:description			
Marine Region	xsd:anyURI	schema:location			
Sampling Platform	xsd:string				
Depth	xsd:integer	qudt:numericValue	http://vocab.nerc.ac.uk/collection/P02/current/AHGT/	metre	http://vocab.nerc.ac.uk/collection/P06/current/ULAA/
Protocol Label	xsd:string	schema:description			
Sampling Campaign	xsd:string	schema:description			
SAMPLE_Objective	xsd:string	schema:description			
EVENT_Device	xsd:string	sosa:madeBySampler			
EVENT_Method	xsd:string	sosa:usedProcedure			
EVENT_Comment	xsd:string	rdfs:comment			
SAMPLE_Quantity	xsd:double	schema:description		litre	http://vocab.nerc.ac.uk/collection/P06/current/ULIT/
SAMPLE_FiltrationTime	xsd:integer	schema:Duration		minutes	http://vocab.nerc.ac.uk/collection/P06/current/UMIN/
SAMPLE_Container	xsd:string	schema:description			
SAMPLE_Content	xsd:string	schema:description			
SAMPLE_Size-Fraction_UpperThreshold	xsd:string	schema:description		micrometre	http://vocab.nerc.ac.uk/collection/P06/current/UMIC/
SAMPLE_Size-Fraction_LowerThreshold	xsd:double	schema:description		micrometre	http://vocab.nerc.ac.uk/collection/P06/current/UMIC/
SAMPLE_Treatment_Chemicals	xsd:string	schema:description			

SAMPLE_Treatment_Storage	xsd:string	schema:description		centigrade	http://vocab.nerc.ac.uk/collection/P06/current/UPAA/
Temperature	xsd:double	qudt:numericValue	http://vocab.nerc.ac.uk/collection/P02/current/TEMP	centigrade	http://vocab.nerc.ac.uk/collection/P06/current/UPAA/
Salinity	xsd:double	qudt:numericValue	http://vocab.nerc.ac.uk/collection/P02/current/PSAL/	PSU	http://vocab.nerc.ac.uk/collection/P06/current/UGKG/
CTD_Conductivity	xsd:double	qudt:numericValue	http://vocab.nerc.ac.uk/collection/P02/current/CNDC/	mS/cm	http://vocab.nerc.ac.uk/collection/P06/current/MSPM/
CTD_Fluorescence	xsd:double	qudt:numericValue	http://vocab.nerc.ac.uk/collection/P02/current/FVLT/	volt	http://vocab.nerc.ac.uk/collection/P06/current/UVLT/
CTD_Depth	xsd:double	qudt:numericValue	http://vocab.nerc.ac.uk/collection/P02/current/AHGT/	metre	http://vocab.nerc.ac.uk/collection/P06/current/ULAA/
Seawater Nutrients Concentration_Ammonium	xsd:double	qudt:numericValue	http://vocab.nerc.ac.uk/collection/P02/current/AMON/	micromol/l	http://vocab.nerc.ac.uk/collection/P06/current/UPOX/
Seawater Nutrients Concentration_Nitrate	xsd:double	qudt:numericValue	http://vocab.nerc.ac.uk/collection/P02/current/NTRA/	micromol/l	http://vocab.nerc.ac.uk/collection/P06/current/UPOX/
Seawater Nutrients Concentration_Nitrite	xsd:double	qudt:numericValue	http://vocab.nerc.ac.uk/collection/P02/current/NTRI/	micromol/l	http://vocab.nerc.ac.uk/collection/P06/current/UPOX/
Seawater Nutrients Concentration_Phosphate	xsd:double	qudt:numericValue	http://vocab.nerc.ac.uk/collection/P02/current/PHOS/	micromol/l	http://vocab.nerc.ac.uk/collection/P06/current/UPOX/
Seawater Nutrients Concentration_Silicate	xsd:double	qudt:numericValue	http://vocab.nerc.ac.uk/collection/P02/current/SLCA/	micromol/l	http://vocab.nerc.ac.uk/collection/P06/current/UPOX/
Seawater Chemical Properties_pH	xsd:double	qudt:numericValue	http://vocab.nerc.ac.uk/collection/P02/current/ALKY/	dimensionless	http://vocab.nerc.ac.uk/collection/P06/current/UUUU/
Seawater Chemical Properties_Dissolved Oxygen concentration	xsd:double	qudt:numericValue	http://vocab.nerc.ac.uk/collection/P02/current/DOCC/	micromol/kg	http://vocab.nerc.ac.uk/collection/P06/current/UPOX/
Seawater Optical Properties_Downward PAR	xsd:double	qudt:numericValue	http://vocab.nerc.ac.uk/collection/P02/current/VSRW/	microE/m^2/s	http://vocab.nerc.ac.uk/collection/P06/current/UM2S/
Seawater Optical Properties_Turbidity	xsd:double	qudt:numericValue	http://vocab.nerc.ac.uk/collection/P02/current/ATTN/	FTU	http://vocab.nerc.ac.uk/collection/P06/current/USTU/
Organic Matter Concentration (Amount or Mass)_DON	xsd:double	qudt:numericValue	http://vocab.nerc.ac.uk/collection/P02/current/TDNT/	mg/Ll	http://vocab.nerc.ac.uk/collection/P06/current/UMGL/
Organic Matter Concentration (Amount or Mass)_POC	xsd:double	qudt:numericValue	http://vocab.nerc.ac.uk/collection/P02/current/CORG/	microg/l	http://vocab.nerc.ac.uk/collection/P06/current/UGPL/
Organic Matter Concentration (Amount or Mass)_PON	xsd:double	qudt:numericValue	http://vocab.nerc.ac.uk/collection/P02/current/NTOT/	microg/l	http://vocab.nerc.ac.uk/collection/P06/current/UPOX/
Organism Concentration (Amount Volume or Mass)_Pigment concentrations	xsd:double	qudt:numericValue	http://vocab.nerc.ac.uk/collection/P02/current/CPWC/	mg/m^3	http://vocab.nerc.ac.uk/collection/P06/current/UMMC/
Organism Concentration (Amount Volume or Mass)_Nano/Microplankton	xsd:double	qudt:numericValue	http://vocab.nerc.ac.uk/collection/P02/current/MATX/	per m^3	http://vocab.nerc.ac.uk/collection/P06/current/UPMM/
ENA_BioSampleID	xsd:string	schema:identifier			
ENA_BioProjectID_TargetGene18S	xsd:string	schema:identifier			

ENA_BioProjectID_TargetGeneMG	xsd:string	schema:identifier			
ENA_BioProjectID_TargetGene16S	xsd:string	schema:identifier			

Genomics data in ENA

The sequences (16S, 18S, and the shotgun metagenomes) derived from the OSD samples are all archived in ENA as fastQ files. All sequences are open access.

The OSD2014 biosample, bioarchive, and sample accession numbers can be found in the data table that is included in the PANGAEA record. We have additionally added the run accession numbers to these in the OSD2014 data we have placed in the OSD github account.

On ENA, the following projects gather the OSD2014 data together

- PRJEB5129 is the umbrella study, and includes
    - PRJEB8682 includes all the amplicon and metagenome samples from the main OSD14 event (467 entries)
    - PRJEB9694 provides access to pre-processed reads where overlapping raw reads from the study PRJEB8682 where merged into a single longer read (150 entries)
- PRJEB22092, is also OSD2014 but is not included in PRJEB5129: the Third-Party Annotation assembly was derived from the primary whole genome shotgun dataset PRJEB8682 (140 entries)

The OSD2018 and 2019 biosample and run accession numbers, together with the sample IDs used in the logsheet-derived data and the sample titles are provided in ENA, are provided as a CSV file on the OSD github account. The following access numbers are used for these years

- OSD2018: project PRJEB40757 and PRJEB40760 hold the 16S and shotgun metagenomes, PRJEB55999 for the 18S
- OSD2019: PRJEB40762 and PRJEB40764 hold the 16S and shotgun metagenomes, PRJEB56005 for the 18S

The 18S sequences were processed twice initially together with the 16S data but it was subsequently realised that a significant improvement in the 18S processing was necessary. The new sequences were uploaded to the project mentioned above and the previous sequences were suppressed.
When uploading the sequences to ENA, the samples were described following the MicroB3 checklist and water samples checklist, with the metadata taken from the logsheets.

Details on how the sequencing for the OSD2018 and 2019 was done will be provided in an Ocean Sampling Day data paper that is being written.

Note that all material samples were destroyed in the process of sequencing them, and no biobanked replicates were retained.

Quality Control

As many of the logsheets were filled in by hand, illegible entries had to be discarded unless corrections were received. Manually transferring the information from the scans of the logsheets to the google sheets was done with extreme care to avoid mistranslations. A check of the coordinates given by the sampling stations using the online LifeWatch Data Service Tool, to ensure that the coordinates lay in the sea or at the coast. A check that the coordinates for the stations were the same in 2014, 2018, and 2019 was also made, as the station IDs used are the same in all years. It is noted that the coordinates of the sampling sites were kept up to the 4th decimal (where so-provided), as several of such sites are coastline spots and this level of precision is necessary to avoid them being placed on land. Finally, an outlier detection was run on the environmental values, and flagged values were double-checked for mistypes or for the use of non-recommended units.

Obtaining 100% compliance with the instructions concerning the collecting and recording of the environmental measurements for the OSD years run by ASSEMBLE Plus proved to be a challenge. Most stations did not specifically indicate the units in which the measurements were made, and here it was assumed that the requested units were used. However, for the 2018 and 2019 logsheets, approximately 40% and 15% of values were recorded as being in different units to those requested (e.g. salinity is given as psu or %; dissolved oxygen concentration as mg/l, ml/l, or %; etc). This plays havoc with the interoperability of the dataset and so had to be dealt with *a-posteriori*. We chose to publish two versions of the 2018 and 2019 logsheet data on the OSD github account:

1. Version 1 includes the values as recorded by each station, including the measurement unit used. Flag columns are used to indicate where the unit given was not that specified in the Handbook.
2. Version 2 includes the conversion of the measurements with the incorrect units to the correct units. Where no conversion was possible, the value was instead discarded and flagged as such. The conversions carried out are documented in GitHub, and were for: downward PAR given in photons/m2/s rather than mE/m2/s; nutrients given in micromol/l instead of g/mol; dissolved oxygen concentration given in mg/l, %,or mol/m3 instead of micromol/kg; turbidity given in NTU rather than FTU (no conversion possible, only FTU retained); dissolved organic nitrogen in moles instead of grams (no conversion possible, data discarded). Flag columns were used to indicate where conversions were made/not possible.

As standardisation of the metadata, environmental measurements, and sequence data from OSD are of vital importance for the outputs from all years to be interoperable and combinable, it was decided not to publish the data from some stations: those providing logsheets with missing or illegible mandatory metadata, and those for which the physical samples were damaged and could not be sequenced, were discarded.

Standard vocabularies used

Sampling, event, and environmental (meta)data

Collected environmental parameters

All of the measured environmental parameters and their units are BODC terms (the NERC vocabulary server). ENVO terms are also included.

Locations, dates, institutes

Dates are given in ISO format (only a start date is included in the IMIS data table, while start and end are included in PANGEA; the duration between the two is minor).

Locations are given as longitude and latitude; also recorded are the Marine Region (MRGid is given), country code (ISO 3-letter standard), and the LME.

IDs, site names

ID are assigned to each OSD site as they joined, as a simple OSD#. These site IDs do not change from year to year. An EventID is assigned to each sampling event following the convention site_date+time, and a MaterialSampleID is assigned following the convention site_date_depth_protocol.

Event and sampling metadata

The MicroB3 recommended sampling parameters are included (e.g. platform, device, protocol). We have linked these individual terms to the closest match in a standard vocabulary where possible (e.g. SAMPLE_Quantity is matched to the volume concept in BODC, IDs are linked to schema.org “identifier”).

The semantics are described in the metadata CSV files on the OSD GitHub account, in the respective OSD year repositories (e.g. the file called OSD18Metadata_MachineReadable_version2.csv for that year and version combination).

Genomics metadata

The metadata uploaded to ENA are taken from the logsheets but do not include the semantic annotations (this is not provided for in ENA). The metadata added to ENA follow the MicroB3 checklist for the 2014 sequences, and the default ENA checklist with additional MicroB3 fields for the 2018, 2019 sequences.

Ongoing work

Additional work is ongoing on the FAIRness of the OSD data and their publication

- The data (environmental measurements, event metadata, and sequence accession numbers) are being formatted into DwC-A to be published on EurOBIS. This will improve **F A I** and **R** for these data.

- All 16S sequences from OSD2014, 2018, and 2019 are being processed and the taxonomic inventories will also be published on EurOBIS. This will improve **F A I** and **R** for these data.
- The OSD GitHub repositories are packaged as Ro-Crates and additional work on improving this is ongoing. This allows for an improved machine accessibility for the entire dataset and not just individual files therein. A minimum of provenance information is provided in the commit comments for all files. This will improve the **Interoperability** of these data.
- To this will be added a more detailed provenance provision for all OSD data in GitHub will be created once the EOSC-Life[1] Common Provenance Model has been incorporated in the Ro-Crate specification (this work will not be completed until just after the ASSEMBLE Plus project ends). This will improve the **RE-usability** of these data.

[1] Horizon 2020 programme grant agreement number 824087