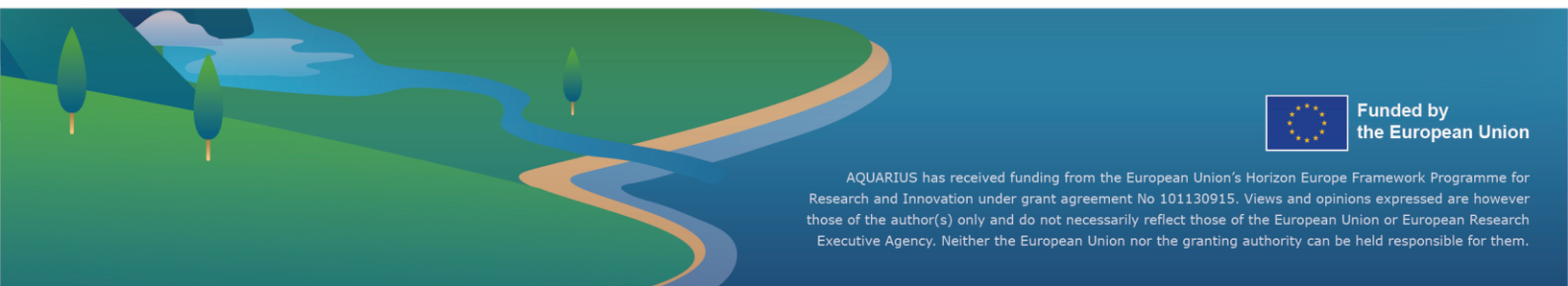




Data Management Plan

DEEP-IMPACT



Funded by
the European Union

AQUARIUS has received funding from the European Union's Horizon Europe Framework Programme for Research and Innovation under grant agreement No 101130915. Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or European Research Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

AQUARIUS DEEP-IMPACT Data Management

1. Data Summary

What is the purpose of the data collection/generation and its relation to the objectives of the project? What types and formats of data will the project generate/collect?

The purpose of the data collection and generation in the DEEP-IMPACT project is to support its overarching objectives, which focus on understanding the dynamics and ecosystem effects of dense-water processes within the Kerkira-Kefalonya valley. Specifically, the data serves to:

1. **Characterize Dense-Water Processes:** Collect data on water temperature, salinity, velocity, and other physical oceanographic parameters to understand the formation, transport, and dissipation of dense water masses in the region.
2. **Analyze Advection Pathways:** Generate and analyze data on pollutant and nutrient transport, mapping how these materials are influenced by dense-water movements.
3. **Assess Ecosystem Impacts:** Gather biological and ecological data to study the influence of dense-water processes on local ecosystems, including benthic communities and pelagic food webs.
4. **Model and Predict Dynamics:** Use data to inform and validate hydrodynamic and biogeochemical models, providing predictive capabilities for dense-water dynamics and their ecological consequences.
5. **Quantify Anthropogenic Influence:** Integrate data on pollutant levels and sources to evaluate the human impact on dense-water processes and associated ecosystems.
6. **Mapping new geographic area:** Collect bathymetric data to produce a more accurate bathymetry of the area.

By aligning data collection and generation with these objectives, the project ensures a comprehensive understanding of how dense-water dynamics affect both natural ecosystems and human-related processes in the study region.

What types and formats of data will the project generate/collect?

The DEEP-IMPACT project is expected to generate and collect a variety of data types in different formats to address its objectives. Here's an overview:

1. Physical Oceanographic Data

- **Types:** Temperature, salinity, pressure, and density profiles. Density is calculated from T, S and P using TEOS-10 equation of density anomaly.
- **Sources:** CTD casts.
- **Formats:** NetCDF

2. Biogeochemical Data

- **Types:** dissolved oxygen concentration.

- **Sources:** Water samples collected during field campaigns and probe measurements
- **Formats:**
 - Excel/CSV (for lab-analyzed data).
 - NetCDF (for probe data, to be included in the Physical data NetCDF).

3. Biological Data

- **Types:** Species composition, biomass, and biodiversity metrics of benthic and pelagic organisms.
- **Sources:** Sediment grabs, ROV (Remotely Operated Vehicle) surveys, and omics.
- **Formats:** CSV/Excel (for species data).

4. Pollutant

- **Types:** Microplastics, heavy metals and other pollutants.
- **Sources:** Seawater, sediment, and biota samples.
- **Formats:** CSV/Excel (for chemical analysis results).

5. Geospatial Data

- **Types:** Bathymetry, sediment type distribution, and habitat mapping.
- **Sources:** Multibeam echosounders, side-scan sonars, and GIS-based analysis.
- **Formats:**
 - Shapefiles/GeoJSON (for vector data).
 - NetCDF (for raster data like bathymetry).

8. Metadata

- **Types:** Logbooks, deployment details, and observational notes.
- **Formats:**
 - CSR (Cruise Summary Report) and PDF (for cruise reports).
 - SeaDataNet standards (for all metadata).

How is the original data gathered on the TA installation(s) and how do you plan to transfer it to you?

As for Transnational Access (TA) data with direct access, these will be taken at the end of the measurement survey campaign.

The results of the analyses performed in the TA labs, on the other hand, will be allocated on the OGS server, as will all the results of the analyses that will be done in the project participants' infrastructures. The data collected in the remote access mode will be sent to the OGS server via internet upload.

What processing on the raw data do you plan? Please differentiate between data quality assurance (handling of outliers, missing and suspect values, null observations) and data harmonisation (code, label and QC flag explanations, consistent use of headers and units, data formatting, conforming to standards).

Processing raw data for the DEEP-IMPACT project involves data quality assurance (QA) and data Quality Control according to international standards, common vocabularies and best practice (e.g. SeaDataNet, OBP) which are essential to ensure the FAIRNESS of the data produced.

Here's a detailed differentiation and plan for both processes:

Physical and biogeochemical oceanographic data

Quality assurance on data collection and Quality Control procedure based on best practice methodologies (e.g., spike test, range test, z-scores, interquartile range) and comparison against historical data or climatologies. Quality Control resulting flags (according to L20 SeaDataNet measurand qualifier flags) will accompany the data.

Biological, pollutant and sediment data

Quality assurance on data collection and Quality Control procedure check based on best practice methodologies (e.g. analysis of variance between different sample replicates).

Geospatial Data

The geospatial data are pre-processed onboard during the campaign and then post-processed following standard procedure and software for multibeam processing.

Multibeam bathymetric data were pre-processed onboard during the survey and will be further post-processed and quality controlled using HYPACK MBMAX and Teledyne PDS. Vessel motion and heading corrections were applied during acquisition using the vessel's inertial navigation system, and sound velocity profiles were collected daily and applied to the acquisition system. During post-processing, soundings will be gridded and visually inspected to identify and remove outliers and erroneous beams. Final bathymetric surfaces and digital terrain models will then be generated in Teledyne PDS.

When do you plan to perform these processing steps by means of a concise planning?

In every phase of data collection will be compiled survey logs with detailed description of used methodology. During the survey preliminary control on the acquired data will be done and at the end of the survey the data processing will be done for physical and biochemical data collected and bathymetric data.

Regarding biological and pollutant data, laboratory analyses may require up to 12 months after the end of the experiment, due to the large number of samples collected. All resulting datasets will nevertheless be finalised, submitted to the relevant data repositories and made available within the lifetime of the AQUARIUS project, allowing sufficient time for validation and publication.

2. Making data FAIR

2.1 Making data findable, including provisions for metadata

*Will you in principle be able to follow the **AQUARIUS Data Management approach** for metadata? What possible challenges do you foresee?*

The dataset generated in the DEEP-IMPACT project will follow the principle of the **AQUARIUS Data Management approach**.

How will you document all this information before submission, especially lineage information (i.e. processing and QC steps)?

The expected size of the total data acquired during the oceanographic campaign is about 5-10 Tb of which approximately 80% of ROV video recording and 15% of multibeam data delivered in grid format. The data acquired will be useful for other projects and outreach activities.

Documenting information, especially lineage (processing and QC steps), is essential for transparency, reproducibility, and compliance with data-sharing standards. Here's how the DEEP-IMPACT project can systematically document this information:

Adopt established metadata standards to ensure thorough documentation of dataset attributes, including lineage.

Metadata Standards

- Metadata makes use of **XML formats**, based upon the **ISO 19115** content model, exchange schema's.
- **CF Metadata Conventions**
- **SeaDataNet metadata formats for:** Physical datasets, Chemical datasets, Biological datasets, Bathymetric datasets, Geological datasets.
- Controlled **common vocabularies** for the marine domain. In the case of SeaDataNet, the vocabularies referred to the NERC Vocabulary Server (NVS 2.0). In the case of omics data we refer to the Genomics Standards Consortium's MIXS checklists and extensions.
(<https://genomicsstandardsconsortium.github.io/mixs/>). It will use the ENA's checklists that comply with the GSC MIXS, most likely the water checklist (<https://www.ebi.ac.uk/ena/browser/view/ERC000024>).

Content to Include in Metadata:

Metadata of the Survey:

Name, coordinating organization (EDMO code), Principal Investigator, Start date, Sea Area (C19 - SEAVOX SALT AND FRESH WATER BODY GAZETTEER), Abstract.

Metadata of the Activity:

Activity name, Organization responsible of the activity, Activity reference person

Metadata of the Dataset:

Dataset Name, Dataset originator, Dataset originator person, Type of device (Instrument class L05 - SEADATANET DEVICE CATEGORIES), Instrument (Instrument code L22 - SEAVOX DEVICE CATALOGUE), Discipline (P08 - SeaDataNet Parameter Disciplines), Parameters group (P03 - SeaDataNet Agreed Parameter Groups), Discovery parameter (List of variables, P02 - SEADATANET PARAMETER DISCOVERY VOCABULARY), Temporal extent, Spatial extent (Dataset Latitude, longitude or spatial domain - bounding box), Coordinate reference system, Data licence (Data Access Restriction, L08 - Data Access Restriction (e.g. CC-BY-4.0)), QC procedures (L12 - SEADATANET PARAMETER QUALITY CONTROL PROCEDURES), Endpoint of Data and Resources (Upload or Link), Resource type (Web service, Geospatial dataset, Map viewer, ...), Persistent Identifiers (DOI).

The data summary log will be in line with the **AQUARIUS TA Data Summary Log**.

QC protocol

The QC protocol involves QC process of anomalous data detection that apply after data gathering:

- Data integrity checks (e.g., format)
- Data value checks (range checks, spikes and outliers checks, neighbor checks, climatology checks)

- SeaDataNet harmonised scheme of QC Flags will be used to label individual data values
- QC Flag scale is the Common Vocabularies L20.

Here an example of workflow

1. Raw Data: Stored raw data files accompanied by survey log describing the collection method.
2. Pre-Processing: Log steps like unit conversions, format changes, and visual inspection of the data (preliminary QC).
3. QC and Processing: Use automated scripts that generate logs of flagged and corrected values. Will be applied mandatory QC test depending on type of data collected (oceanographic, biological etc.) and resulting QC flags will accompany the data.
4. Final Dataset: Save the final processed dataset in the appropriate format depending on the variable (NetCDF, Excel/.csv, Shapefiles/GeoJSON etc) alongside with separate files for:
 - QC summary report.
 - Provenance file detailing the workflow.

After the generation, the dataset will be submitted through the EMODnet Data Ingestion or SeaDataNet SeaNoe service and transferred to expert data centres for inclusion in local data centres and subsequent integration into the applicable European data management infrastructure(s).

File names will follow a consistent and descriptive convention based on the project name (DEEP-IMPACT), data category or dataset content, processing level where applicable, and version number. The exact structure and file format may differ among data types, depending on the nature of the dataset and the standards adopted for its final publication. Examples may include DEEP-IMPACT_OceanographicData_L2_v01.nc or DEEP-IMPACT_BiologicalData_v01.csv.

2.2. Making data openly accessible

Will you collect/generate data that might require a temporary embargo? Indicate which type of data and motivate its temporary embargo for each case.

Do you also plan to make the data and metadata available on another repository next to the AQUARIUS repositories, for instance an institutional, national or general data repository?

No embargo is foreseen for the data collected during the DEEP-IMPACT project for preparing scientific publications or seeking patents and all the data acquired in the DEEP-IMPACT project will be associated with persistent identifiers (DOI) and licensed in CC-BY-4.0. The DEEP-IMPACT project plans to share a metadata repository for the DEEP-IMPACT project also in the National Oceanographic data centre of the OGS.

2.3. Making data interoperable

*Will you in principle be able to organise your TA team during the TA activities in order to gather the input required for the **AQUARIUS TA Data Summary Log**? What possible challenges do you foresee?*

Yes, there is no problem meeting the project specifications and within the time frame.

2.4. Increase data re-use (through clarifying licences)

Do you foresee any difficulty for your collected/generated data to follow this policy?

All metadata and data acquired in the DEEP-IMPACT project will be deposited and published as openly accessible metadata and data in trusted repositories within the lifetime of the AQUARIUS project, under the CC BY 4.0 licence.

3. Other research outputs

Do you plan to generate other research outputs next to the collected/generated data such as digital (e.g. software, workflows, protocols, models, etc.) or physical (e.g. new materials, antibodies, reagents, samples, etc.) results?

No, the DEEP-IMPACT project does not generate other research output.

4. Allocation of resources

The AQUARIUS TA Data Summary Log using the proper app is compiled during TA activity when internet connection works properly. The Cruise Summary Report will be completed within the defined deadline (one month after the end of the project). The original metadata and data will be transformed to the AQUARIUS prescribed standards, performing QA-QC, and making these FAIR and ready for inclusion in AQUARIUS repositories, thereby supported by AQUARIUS expert data centres. The responsible of Data Management is the P.I. of the project.

5. Data security

What provisions are in place by the TA teams for data security (including backups, secure storage and transfer) in the initial phase, before the data is transferred to the AQUARIUS expert data centres?

Data will be transferred to OGS servers that provide all necessary security features. In addition, data will be transferred to personal portable memories to provide additional security, also during the survey the data will be backuped every day after the sampling.

6. Ethical aspects

Are there, or could there be, any ethics or legal issues that can have an impact on data sharing?

The DEEP-IMPACT project does not include ethical aspects or legal issues to be taken into consideration.

7. Other issues

Will you follow any other national/ sectoral/ departmental procedures for data management, on top of the infrastructure and procedures that AQUARIUS will provide as part of its data management approach?

The DEEP-IMPACT project will follow the procedures that AQUARIUS will provide as part of its data management approach.