AQUARIUS Final Data Management Plan for TA Call applications

November 2025

1. Data Summary

➔ What is the purpose of the data collection/generation and its relation to the objectives of the project?

The data collection aims to characterize dynamics of air-sea VOC exchange and shipping emissions in the Baltic Sea region. This aligns with the project's objectives to understand the abundance and diversity of marine VOCs, as well as their role in assessing the impact of shipping activities on local air quality. The data will provide insights into both natural marine processes and anthropogenic influences on atmospheric composition.

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

Data collected during the project consists of data collected continuously at Utö by the facility provider. This data include a wide range of atmospheric observations like wind speed and direction, global radiation, T and RH, turbulence, air quality parameters (e.g. PM10, SO2, NOx, O3), biogeochemical variables (like T, S, O2, pH, Chla, turbidity, phytoplankton community composition etc) measured from the sea water with a flow-through system.

In addition to these existing and additional measurements provided by the facility owner, TA user group will bring a new instrument to the site. The instrument (PTR-TOF-MS) is listed in the application. It provides high-frequency (10 Hz) mass spectra of VOC counts-per-second (cps) as raw files in a HDF5 (Hierarchical Data) format (instrument native). Raw acquisition uses vendor software for control/diagnostics (IoniTOF 4; will be recorded in the metadata) but this software is not required to read the raw HDF5 files. Files are saved in 30-minute intervals (~18,000 records per file) and include continuous measurement (the main data), daily 30-minute "ZeroAir" instrument backgrounds, and pre/mid/post calibration files. These flags will all be recorded in the metadata. Preprocessing of the raw files will be performed with PTRwid (Holzinger, 2015), a free tool implemented in IDL and runnable under the freely available IDL Virtual Machine[1]. PTRwid produces per-file CSV tables using a unified mass list (rows = timestamps, columns = identified mass peaks); these CSVs plus the raw HDF5 files will be published to ensure accessibility.

For long-term access and reuse we will also provide mixing-ratio time series in the same format as PTRwid output (CSV tables), together with metadata describing variables, units, QA/QC flags, and formulae for converting counts to mixing ratios. All conversion and processing scripts (Python) and example notebooks will be version-controlled in a public repository and archived with a DOI; the metadata record will include instructions for reproducing conversions without proprietary tools. Following the mixing-ratio derivation, we will compute eddy-covariance fluxes for each identified VOC using the free EddyPro software[2]. We will publish, for each mass (compound), the resulting flux time series, and the full set of EddyPro QC outputs/flags and diagnostics required for flux evaluation and screening.

Regarding the observations by the facility provider, the operational copies are stored first on FMI servers in ascii format and shared after the quality control through FMI open-access channels (Utö station webpage[3], FMI data portal[4], itämer.fi) for rapid

[1] https://www.nv5geospatialsoftware.com/Support/Maintenance-Detail/the-idl-virtual-machine; The IDL Virtual Machine is free to use and does not require a license to run. However, it is required to register for an account on the NV5 Geospatial Software website in order to download the installers.

[2] https://www.licor.com/products/eddy-covariance/eddypro .

[3] https://swell.fmi.fi/Uto/

[4] https://en.ilmatieteenlaitos.fi/download-observations

access. The authoritative, curated versions will be deposited in SeaDataNet (Primary) and ICOS Research Infrastructure, from which EMODnet will harvest; FMI portals and itämeri.fi act as Secondary exposure channels for visibility and regional users. FMI portals and itämeri.fi act as Secondary exposure channels for visibility and regional users.

The table below details of the different datasets based on SDN common vocabularies (if available):

Instrument Type [L05]	Instrument Model [L22]	Measured/Processed variables [P01]	File format(s)	Primary repository	Other repository	Expected size
proton transfer reaction mass spectrometers [LAB46]	Ionicon PTR-TOF-MS series proton transfer reaction mass spectrometer [**not found**]	Raw spectra [**not found**]	*.h5	SeaDataNet	ERDA (University of Copenhagen)	~6 Terabytes
		processed count-per-second (with PTRwid) [**not found**]	*.csv	SeaDataNet	ERDA	~6 Terabytes
		VOC mixing ratios (with Python) [**not found**]	*.csv	SeaDataNet	ERDA	~1 Terabyte
		VOC fluxes using Eddy Covariance (with EddyPro) [**not found**]	*.csv	SeaDataNet	ERDA	~300 Gigabytes

Table 1. Datasets information

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

Files are copied from the acquisition laptop to external hard drives using Total Commander with "Verify after copy" enabled. This provides a byte-by-byte verification immediately after each copy. For the first half of the campaign, data

integrity was confirmed through successful preprocessing in PTRwid (all files opened and processed without errors). This provided practical assurance that files were intact and readable. For the second half of the campaign (from July 2025), we generate SHA-256 checksum files both on the PTR acquisition laptop (source) and on the transporting hard drives (destination), and we confirm the checksums on receipt. This provides end-to-end integrity verification. The original on-site copy is retained until off-site verification is complete. Furthermore, Utö station has both wireless and fixed high-speed network and the island is connected to the mainland through optical fibers. If needed, individual files may be transferred over the network (e.g., via RealVNC-mediated access) for rapid diagnostics; such transfers are verified by opening the files in PTRwid or with appropriate Python libraries.

➔ 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).

All data producers (facility provider and TA user groups) apply quality assurance and quality control procedures following community-standard protocols. Quality assurance includes adherence to standard operating procedures (SOPs), manufacturer calibration schedules, systematic logging of instrument metadata (calibration dates, zero-air/background measurements, changes to instrument parameters), and standardization of all timestamps to UTC.

The processing steps for PTR-TOF-MS data are detailed in the previous question. For its quality control, we use a staged flagging system tailored to each processing step. Raw data enter the preprocessing pipeline flagged only by measurement type (sample/background/calibration). The resulting count-per-second time series are checked for drift or sudden changes in instrument parameters and primary ion counts. For mixing ratios, values below the limit of detection (LOD), outliers, and suspect values are flagged rather than removed wherever possible to preserve data traceability and allow reprocessing. Computed fluxes follow the (Mauder and Foken 2004) convention (three-tier flagging system). Missing data and null observations are explicitly marked with standardized flags and documented in metadata.

Data harmonization: Variable names, units, and QC flag definitions are documented in standardized metadata accompanying each dataset. For VOCs measured by PTR-TOF-MS, we use IUPAC chemical names and CAS registry numbers, with cross-references to ChEBI/PubChem identifiers where applicable. Only VOCs present in the calibration gas mixture can be identified and labeled; uncalibrated mass peaks are reported as "m/z + 1" values representing their protonated mass-to-charge ratio. When tentative compound assignments are made to uncalibrated masses (e.g., based on literature or mass spectral interpretation), the standard IUPAC names and identifiers mentioned above will be used, and the assignment will be flagged as tentative in the metadata.

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

The PTR-TOF-MS instrument will return to Copenhagen in mid-October 2025, along with the final batch of data on external hard drives. Processing the raw data in PTRwid requires loading all data simultaneously to generate a unified mass list covering the entire measurement period. Based on previous experience, this initial preprocessing step takes up to 45 days. Including subsequent QC, calibration application, and flux computation steps, we anticipate completing all the processing by the end of January 2026.

➔ What is the expected size (Megabyte to Terabyte range) of the data? Please note that in case of multibeam or imagery data, the derived products (such as grids) should be submitted instead of the raw measurements.

The expected size for each output is outlined in Table 1.

➔ To whom might your data be useful, outside your project?

The data will be valuable to multiple research communities: (1) ocean biogeochemistry researchers studying biogenic volatile organic compound production and marine carbon cycling, (2) atmospheric chemists investigating air-sea exchange processes and aerosol formation, (3) climate modelers needing high-resolution VOC flux parameterizations, and (4) shipping emissions researchers quantifying anthropogenic impacts in coastal zones. As one of the first applications of high-frequency PTR-TOF-MS in marine settings, the dataset also provides a methodological reference for researchers planning similar deployments. While the untargeted analysis framework (tens of VOCs at 10 Hz) hasn't yet been integrated into standardized marine data pipelines, this dataset can help establish those standards and demonstrate the potential for routine monitoring applications.

2. Making data FAIR

2.1 Making data findable, including provisions for metadata

The **AQUARIUS Data Management approach** is aimed at inclusion of your collected/generated data in leading European aquatic data management infrastructures such as **SeaDataNet**[5] (physics, chemistry, geophysics, geology, and biology), **EurOBIS**[6] (marine biodiversity), **ELIXIR-ENA**[7] (biogenomics), **ICOS-Ocean**[8] (carbon), and **Copernicus INSTAC**[9] (Near- Real-Time data). Inclusion will be mediated through expert data centres which are partners in the AQUARIUS project. They will give training and education to the TA teams about planned data management practices as part of AQUARIUS WP5. In addition, as part of AQUARIUS WP6, they will give support to the TA teams for the adoption of the preferred metadata and data standards, and ingestion of the resulting metadata and data in repositories that are nodes of the indicated European data management infrastructures. This way, the metadata and data from the TA projects will become FAIR and also available for EMODnet, Copernicus Marine, DTO developments, EOSC, and global data exchanges such as GEOSS and the digital ocean ecosystem which is being developed in the framework of the UN-IOC-Ocean Decade programme. In the AQUARIUS Data Management approach, it is recommended to make use of the online **Data Submission service of EMODnet Data Ingestion**[10] and/or **SeaDataNet SeaNoe service**[11] for transfer of your metadata and data packages to the AQUARIUS expert data centres.

Paragraph 5.4 of the **AQUARIUS Data Management approach** document lists the minimum set of metadata that EMODnet Ingestion requests you to enter in order to document your data packages submission. In addition, Annex 1 of the document gives a list of items that should be documented during the TA data acquisition and processing activities

[5] https://www.seadatanet.org

[6] https://www.eurobis.org

[7] https://www.ebi.ac.uk/ena/

[8] https://www.icos-cp.eu/observations/ocean/otc

[9] https://marine.copernicus.eu/about/producers/insitu-tac

[10] https://www.emodnet-ingestion.eu

[11] https://www.seadatanet.org/Software/SEANOE

as these are later needed, when transforming data to the standard metadata and data formats as required.

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

  Yes. We will complete all required EMODnet Ingestion metadata fields (section 5.4) and maintain comprehensive documentation following Annex 1 recommendations, including PTR-TOF-MS instrument specifications, calibration records, processing scripts, and QC procedures. We will use the AQUARIUS TA Data Summary Log systematically throughout the campaign. Note that only one observation entry was recorded as the PTR-TOF-MS operated as a continuous measurement system.

  Main challenge: As PTR-TOF-MS high-frequency VOC measurements are an emerging marine data type, standardized metadata formats are still developing. We will work with AQUARIUS data experts to define appropriate schemas that maintain EMODnet compatibility while capturing all instrument-specific information for data re-use.

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

  We will maintain detailed electronic logbooks documenting all processing steps, calibrations, and QC procedures. This will include:

  - Daily instrument performance logs
  - Calibration records and standards used
  - Processing scripts with version control
  - QC flagging criteria and decisions

2.2. Making data openly accessible

The AQUARIUS project follows an **open data access policy.** Therefore, the **AQUARIUS Data Management approach** is aimed at inclusion and publishing of your collected/generated data as open metadata and data in trusted repositories, within the lifetime of the AQUARIUS project. However, temporary embargos (max 2 years after collection) are possible for specific data that require additional processing and to allow for scientific analyses that will lead to scientific papers.

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

  The collected dataset is unique and the scientists need time to publish the results in scientific journals. A short embargo is important especially in the case of PhD students using the collected data. This embargo is only for BVOC data collected during the project - most of the data collected at Utö is openly available. Typically, the main limiting factor for providing open-access data is the availability of labour to do the QA/QC for the non-standard data.

- ➔ 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?

  Yes, in addition to the primary AQUARIUS repository (SEANOE.org via SeaDataNet), we will store data in University of Copenhagen's ERDA system for institutional backup and working access. We will ensure proper linkage between repositories using DOIs to maintain traceability and avoid duplication. The SEANOE DOI will be cited in any institutional records and publications.

➔ In case of use of exotic data formats for original data, what methods or software tools are needed to read and understand the data and could you provide documentation about that (including possible open-source code)?

This is already addressed in section 1. Data Summary

2.3. Making data interoperable

The **AQUARIUS Data Management approach** is aimed at making the collected/generated data interoperable by adopting common standards from leading European data management infrastructures. Use will be made of standard formats, QA-QC procedures, and controlled vocabularies for many metadata tags. For this process the TA teams will be given support by the AQUARIUS expert data centres, who also can help with use of common tools and services. To be able to do these transformations from original metadata and data to standardised metadata and data, it is important that relevant information about the data collection/generation is noted down during the actual events.

This consists of the **AQUARIUS TA Data Summary Log** which should be maintained by the PI of the TA project scientific team to keep an overview and index of the data collection events. It should be shared afterwards with the AQUARIUS data centre experts as it will provide an index to the collected/generated data and relevant documentation, such as who, when, where, how, etc. It contains only metadata and no data. For keeping this log, the PIs of the TA projects will receive a tablet from WP6 with a preloaded app, that will facilitate entering, browsing, and exporting information. The tablet should be returned to the WP6 team after finalisation of the TA acquisition activities. The **AQUARIUS TA Data Summary Log** will serve as a list for the expert data centres to know what data to expect from where and who and as a checklist for the following steps in the data management workflow.

In addition to the TA Data Summary Log it is recommended that the researchers of the TA teams document several aspects of the data acquisition and following processing, so that the later transformation to the common standards of the European marine data management infrastructures will be easier. Annex 1 of the AQUARIUS Data Management approach document gives a checklist which documentation TA researchers are recommended to maintain for this purpose.

➔ 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?

Our measurement campaign involved continuous fixed measurements with a single instrument over an extended period. Consequently, we did not generate an extensive AQUARIUS TA Data Summary Log. The user group leader submitted the log through the app following the campaign's conclusion in October. No significant challenges were encountered in organizing the TA team for this purpose, given the straightforward nature of our measurement setup.

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

The AQUARIUS project follows an **open data access policy,** and the AQUARIUS Data Management approach is aimed at inclusion and publishing of your collected/generated data as open metadata and data in trusted repositories, within the lifetime of the AQUARIUS project. The preferred licence for use of AQUARIUS data is CC-BY-4.0.

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

We are committed to open science principles, and all planned measurements are compatible with open data sharing. However, the main challenge will be the available resources for doing this work, so some delays are expected.

➔ How are you going to organise gathering the input required for the AQUARIUS TA Data Summary Log during the TA activities?
Given that our campaign involves continuous fixed measurements with a single instrument, the logging process was not extensive and the user group leader coordinated the logging. Key metadata, measurement parameters, and any relevant operational notes are documented during the campaign.

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.

➔ If an embargo is sought to give time to publish or seek patents, specify why and how long this will apply, bearing in mind that research data should be made available as soon as possible.

A temporary embargo is requested to allow time for scientific publication of research findings derived from these data. Due to the substantial volume of data generated during the campaign and the lengthy preprocessing and quality control procedures required, the embargo period needs to extend until at least July 2026. This timeframe will enable the research team to complete the necessary data processing, conduct thorough analysis, and prepare manuscripts for peer-reviewed publication. Following this embargo period, all data will be made openly available through the designated repositories. As the produced data is novel in terms of standardization, the user group leader is willing to collaborate with the AQUARIUS data management team to establish appropriate standards for this dataset, ensuring its accessibility and interoperability for the broader scientific community.

4. Allocation of resources

Terms & Conditions of AQUARIUS Funding require TA teams to follow the AQUARIUS Data Management Approach. This includes efforts for:

- updating their initial **Data Management Plan (DMP)** per TA project, as included in their original proposal submission, after being awarded and getting ready for the actual TA project deployment
- taking part in WP5 data management training activities to learn about AQUARIUS data management standards and getting experience with tools and services
- using the app during TA activities to complete the **AQUARIUS TA Data Summary Log**
- completing a **Cruise Summary Report (CSR)**[12] in case the TA project will involve scientific cruises with Research Vessels
- transforming the original metadata and data to the AQUARIUS prescribed standards, performing QA-QC, and making these FAIR and ready for inclusion in AQUARIUS

[12]https://www.seadatanet.org/Metadata/CSR-Cruises

repositories, thereby supported by AQUARIUS expert data centres.

5. Data security

The operators of the repositories as used for the AQUARIUS archival and long-term stewardship have security provisions in use such as for backups, secure storage, firewalls, etc.

➔ 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?

FMI servers storing the Utö data are back-upped. Manual hard-drive copies are taken regularly from the non-network instruments during the field period.

➔ Who will be responsible for data management in your TA project?

User group leader.

6. Ethical aspects

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

No.

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?

User group and facility providers will follow their internal, institutional guidelines on data sharing procedures, which all are in line with AQUARIUS approach.

8. References

Holzinger, R., 2015. PTRwid: A new widget tool for processing PTR-TOF-MS data. Atmospheric Measurement Techniques 8, 3903–3922. https://doi.org/10.5194/amt-8-3903-2015