www.aquacosm.eu

DATA MANAGEMENT PLAN VERSION 1

Deliverable No: D4.4

PROJECT TITLE:

AQUACOSM-plus

Network of Leading Ecosystem Scale Experimental Aquatic Mesocosm Facilities Connecting Rivers, Lakes, Estuaries and Oceans in Europe and Beyond

PROJECT NUMBER:

871081

PROJECT TYPE:

Research and Innovation Action

WORK PROGRAM TOPICS ADDRESSED:

H2020-INFRAIA-2018-2020 / H2020-INFRAIA-2019-1

Deliverable title	**DATA MANAGEMENT PLAN VERSION 1**
Deliverable number	D4.4.
Deliverable version	1
Contractual date of delivery	30 September 2020 (M9)
Actual date of delivery	30 September 2020 (M9)
Dissemination level	Public
Nature of deliverable	Report
Work package	WP4
Lead Beneficiary	NIOO
Author(s)	L.N. de Senerpont Domis, Simon Keeble, Johan Wikner, Vivi Pitta, Ioulia Santi
Editor	
EC Project Officer	Blagovesta CHOLOVA

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 871081

Table of Content

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 871081

1. Abstract

This deliverable provides the first version of the Data Management Plan (DMP) of AQUACOSM-plus. AQUACOSM-plus will collect aquatic mesocosm data from 28 partner institutions hosting > 60 mesocosm facilities throughout Europe. One of the most central aims of AQUACOSM-plus is to ensure a self-promoting and lasting management of data flow produced by all involved facilities. As part of the H2020 Open Research Data Pilot a Data Management Plan is developed. The development of the Data Management Plan has followed a stepped approach. This Data Management Plans sets the initial guidelines for how data will be generated in a standardized manner, and how data and associated metadata will be made accessible. This Data Management Plan is a living document and will be updated through the lifecycle of the project.

This Data Management Plan builds on the second version of the DMP of AQUACOSM (project ID 731065), the predecessor of AQUACOSM-plus and has been amended for new developments and initiatives in AQUACOSM-plus. AQUACOSM-plus will introduce a web-interface for data entry, enabling near-real-time data flow directly from the experimental facilities to the databases (WP4, WP5). Complemented by an improved metadata portal (www.aquacosm.eu), open tools for data analyses and processing, and easy allocation of DOI's to datasets, AQUACOSM-plus will create a sustainable Open Science workflow.

2. Data Management Plan

2.1 Definitions, acronyms and abbreviations

BODC: British Oceanographic Data Centre

Copernicus (AKA **CMEMS** – Copernicus Marine Environmental Monitoring Service): a European Union Programme aimed at developing European information services based on satellite Earth Observation and sampling in the field

DMP: Data Management Plan

DOI®: Digital Object Identifier is a persistent identifier used to uniquely identify objects, standardized by the ISO

EML: Ecological Metadata Language

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 871081

FAIR: Research data that is findable, accessible, interoperable and re-usable. These principles precede implementation choices and do not necessarily suggest any specific technology, standard, or implementation-solution.

GB: Gigabytes

GitHub: is a web-based Git or version control repository for source code. A Git is a version control system (VCS) for tracking changes in computer files and coordinating work on those files among multiple people.

ISO: International Organisation for Standardisation, is an international standard-setting body composed of representatives from various national standards organizations.

Metadata: data that provides information about other data. Three types of metadata can be distinguished, including descriptive metadata, structural metadata and administrative metadata.

NERC: Natural Environment Research Council, the United Kingdom's leading public funder of environmental science.

Open data: Research data that can be freely used, re-used and redistributed by anyone for any purpose. Open data is free of restrictions from copyright, patents or other mechanisms of control.

PhD: doctoral degree awarded by universities.

Processed data: also known as secondary data. This data that has been part of a processing routine, "cleaning" by researchers to remove outliers, obvious instrument reading errors or data entry errors, or any analysis (e.g., determining central tendency aspects such as the average or median result). In addition, this data may have been subjected to more statistical forms of analysis

QA: Quality Assurance

QC: Quality Control

R: an open source programming language and software environment for statistical computing and graphics that is supported by the R Foundation for Statistical Computing

Raw data: also known as primary data, is data (e.g., numbers, instrument readings, figures, etc.) collected from a source. True raw data has not been subjected to processing or any other manipulation by a software program or a human researcher, analyst or technician.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 871081
Responsibility for the information and views set out in this report lies entirely with the authors.
The European Commission is not responsible for any use that may be made of the information it contains.

RDA: Research Data Alliance, is a research community organization started in 2013 by the European Commission, the American National Science Foundation and National Institute of Standards and Technology, and the Australian Department of Innovation. Its goal is to build social and technical infrastructure to enable the open sharing of data

SOP: Standard Operating Procedure

TA: Transnational Access. Transnational Access means free of charge, trans-national access to research infrastructures or installations for selected user groups. The access includes the logistical, technological and scientific support and the specific training that is usually provided to external researchers using the infrastructure.

WP: Work package

2.2 Data summary

AQUACOSM-plus will collect aquatic mesocosm data from from 28 partner institutions hosting > 60 mesocosm facilities throughout Europe (see www.aquacosm.eu). AQUACOSM-plus is organised in nine work packages (Table 1).

Table 1: Work packages in AQUACOSM-plus, acronyms of beneficiaries as detailed in the grant agreement of AQUACOSM-plus

WP number	WP title	Lead beneficiary
WP1	Consortium Management	FVB-IGB
WP2	NA1: Science and innovation strategy for society	NORCE
WP3	NA2: Networking and Training for Knowledge Transfer (NTKT)	METU
WP4	NA3: Breaching barriers to open mesocosm science, including open science tools and data	NIOO-KNAW
WP5	NA4: Outreach activities: Communication, Dissemination and Exploitation	HCMR
WP6	NA5: Defining Grand Challenges in aquatic mesocosm research	WCL
WP7	JRA1: Towards transformative mesocosm research – breaking the spatial and temporal barriers of aquatic ecosystem experimentationecosystem studies in all climates	SYKE

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 871081

WP8	JRA2: Pilot execution of Grand Challenge scenario-testing through bridging scales of experimental and observational RI networks	LMU
WP9	TA: Provision of transnational access to all AQUACOSM-plus facilities	FVB-IGB

One of the most central aims of AQUACOSM-plus is to ensure a self-promoting and lasting management of data flow produced by all involved facilities. Building on AQUACOSM, AQUACOSM-plus will introduce a web-interface for data entry, enabling near-real-time data flow directly from the experimental facilities to the databases (WP4, WP5). Complemented by improvements to the existing metadata portal (D4.6 in AQUACOSM 731065; www.aquacosm.eu), open tools for data analyses and processing (D4.2/D4.5/D4.11), a pipeline for allocating DOI's to datasets (D4.8), AQUACOSM-plus will create a sustainable open science workflow.

All experimental data will be collected under WP9, where Transnational Access to all AQUACOSM-plus facilities is realized. With a requested number of 11,500 Transnational Access days, and most TA teams consisting of 3-4 people, staying anywhere between 21-90 days, we anticipate 100-180 datasets to be collected under AQUACOSM-plus (*cf.* 105 datasets in AQUACOSM d.d. 9-9-2020). The expected size of each dataset is relatively small, (<20 GB per experiment) and will contain different types of data, including numeric data, image data, video data, as well as text data. To reduce heterogeneity in data collection and processing, WP4 (D4.2) will develop a common strategy on best practices in reducing (meta) data heterogeneity and processes, including metadata-controlled vocabulary, data format interoperability, as well as file naming conventions. This strategy will be captured in the next version of the DMP to ensure the metadata files and primary data files are quality assured and controlled with long-term value and can smoothly be adapted into a data portal. To ensure uptake of this strategy, WP4 will provide training videos and guidance documentation for TA users (D4.3). It is the goal of AQUACOSM-plus to have all data produced during the lifetime of the project to be findable, accessible, interoperable and re-usable (FAIR). AQUACOSM-plus data will not only be useful for the current and future generation of mesocosm scientists, but also environmental assessment agencies, water quality managers, and companies with a vested interest in water quality.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 871081
Responsibility for the information and views set out in this report lies entirely with the authors.
The European Commission is not responsible for any use that may be made of the information it contains.

2.3 FAIR data

2.3.1 Making data findable, including provisions for metadata

The data produced through TA during the lifecycle of the project will be discoverable via a centralised metadata database. This database has been built as part of the platform of the AQUACOSM web portal, and will later be transferred to the mesocosm.eu website, the one stop portal for mesocosm science. Metadata can be completed online and is currently built upon the Ecological Metadata Language (EML, see Fegraus et al. 2005), with options to extend if required to accommodate future ISO standards. Apart from details on the mesocosm experimental design and methodology, the metadata also contains the contact details of the data owner as well as the Digital Object Identifier (DOI), if available. We will maintain an active connection with the Research Data Alliance (RDA) to follow up on standards used there. A user survey will be carried out on the existing AQUACOSM metadata portal and the results will be used to further improve the portal.

Upon completion of experiments carried out under AQUACOSM and AQUACOSM-plus, metadata should be entered in the metadata database, as part of the Transnational Access requirements. To facilitate that the AQUACOSM-plus DMP is widely embraced in the TA community, a summary of this DMP is provided to each TA user and provider. In addition, a guideline and online videos with instructions on how to use the metadata web filing system will be embedded in the AQUACOSM webportal (www.aquacosm.eu) and provided to both TA user and provider. The AQUACOSM-plus metadata interface is publicly exposed via an API and we will contact relevant data gateways (Pangaea, EMODnet etc.) to encourage them to harvest the metadata and, reciprocally, the AQUACOSM-plus metadata portal will be scheduled to collect other metadata of interest from other portals.

2.3.2 Making data openly accessible

Apart from metadata, as per the grant agreement, primary data should be openly accessible within 6 months after completion of the publishable dataset, with reasons given if this is not possible. The publishable data set is defined as a dataset that has been subject to processing routines aimed at e.g. QA and QC. Reasons for not making the publishable dataset openly accessible may include competitive advantages such as the completion of a PhD thesis, in which case an embargo of up to three years would be accepted.

A node will be set up in Pangaea (with automated uploads to Dataone, Dryad, and other repositories) where mesocosm data may be deposited for easy collection and reuse. All

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 871081
Responsibility for the information and views set out in this report lies entirely with the authors.
The European Commission is not responsible for any use that may be made of the information it contains.

datasets should have a Digital Object Identifier, and we will set up a workflow for allocating DOI's to TA datasets within a European open data portal such as PANGAEA or DRYAD (D4.8). For guidance, AQUACOSM-plus' predecessor AQUACOSM has published guidelines for database management, including controlled vocabulary, which are publicly available through the web portal.

In addition, to pave the way for a future centralised mesocosm open data repository, an application platform for primary data collection will be developed, docked with an existing database and support mesocosm variables and procedures (WP4, task 4.3). This platform will enable capturing of primary mesocosm data, with potential wider application to all experimental data. Data storage tables will be modified to handle experiments, treatments, levels and replicates for providing proper metadata. Web forms for scientific variables will be adapted to requirements by project-partners, for entering primary data. Partner tests will be conducted to identify the necessary modifications. A pilot of the platform will be installed on a server hosted by NIOO-KNAW. User access for entering data will be provided by a web interface and strategies for extracting data from existing databases suggested. Open-access status of the data will be achieved through liaison with existing European data portals, such as Pangaea and EMODnet. By month 48 of the AQUACOSM-plus project it is anticipated that a minimum of 30% of the project partners (100% of those involved in this initial pilot) will have successfully adopted the new platform and a plan to migrate the remaining partners will be provided.

2.3.3 Making data interoperable

Interoperability of data collected within the AQUACOSM-plus life cycle is promoted through the development of a Wikibook on mesocosm science (WP3), through a strategy on reducing heterogeneity in data collection and processing (WP4) and through training and education on various aspects of data collection and processing (WP4/WP5)

To allow for standardised data collection we (WP3, task 3.1) will update and expand the content already collected as best practice guidelines and Standard Operation protocols during AQUACOSM, and produce easily accessible, updatable and practical ***Wiki book.*** This Wiki-based interactive book will provide the state-of-the-art of current knowledge and know-how as well as practical experiences and best practice advices on designing, managing, operating and conducting research at mesocosm infrastructures. The Wiki book will also include guidance on data analyses, in an integrated manner, from freshwater and marine domains including newly developed technologies, and prototypes in mesocosms research across climate zones.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 871081
Responsibility for the information and views set out in this report lies entirely with the authors.
The European Commission is not responsible for any use that may be made of the information it contains.

In addition, by adopting the Wiki 'philosophy', it will enable AQUACSOM to engage and collaborate with mesocosm activities both inside and outside of the partner network.

To reduce heterogeneity in data collection and processing we (WP4, task 4.2) will develop a common strategy on best practices in reducing (meta) data heterogeneity and processes (e.g. metadata-controlled vocabulary, data format interoperability, file naming conventions etc.) to ensure the metadata files and data files are quality assured with long-term value and can smoothly be adapted into a data portal.

We (WP4 Task 4.2 in collaboration with WP3, task 3.2) will provide TA users with training on reducing heterogeneity before each experiment (via webinar / video and online documentation) (in close cooperation with WP3.2 summer schools and workshops for TAs)). During the webinar / video, TA users will receive help and guidance concerning how and where they should upload data (Task 4.3, leading to long-term curation in Task 4.1) and metadata (Task 4.4). During the virtual instructions, TA users will be able to ask questions concerning the procedure.

2.3.4 Increase data re-use (through clarifying licences)

Within the lifecycle of AQUACOSM-plus the data made openly available will be licenced following the service and licence commitment of Copernicus (http://marine.copernicus.eu/services-portfolio/service-commitments-and-licence/). Data collected under AQUACOSM-plus will be made available for re-use upon completion of the experiment. As noted above, a node will be set up in Pangaea (with automated uploads to Dataone, Dryad, and other repositories) where mesocosm data may be deposited for easy collection and reuse. The reuse of data will be promoted by running a Virtual Access pilot using near real-time data streams.

As also noted above, for reasons of competitive advantages a data embargo may apply, including the completion of a PhD thesis, in which case an embargo of three years will be upheld. Data produced and made openly available under AQUACOSM-plus will be available for third parties. Restriction on use of data, software and code are documented in the grant agreement, and may vary according to institutional and national policies and legislations. In case of restrictions on use, metadata is freely still provided, which allow for contacting of the data owner. The request will then be up for consideration of the data owner, and depending on the data owner's decisions full access to the data may be granted. It is the intention to keep the data available indefinitely but additional costs for keeping the web links alive might be applicable.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 871081
Responsibility for the information and views set out in this report lies entirely with the authors.
The European Commission is not responsible for any use that may be made of the information it contains.

3. Dissemination Activities Related to this Deliverable

To facilitate that the AQUACOSM-plus DMP is widely embraced in the TA community, a summary of this DMP is provided to each TA user and provider. In addition, a guideline on how to use the metadata webifying system, embedded in the AQUACOSM web portal (www.aquacosm.eu) is provided to both TA user and provider. In addition, WP4 will provide instructional videos to TA users. Wiki pages will be disseminated via networks and social media, to encourage greater knowledge and collaboration on documentation and standards development. A digital news article will be produced and disseminated via the project website to summarise this information and provide easy access to the documentation.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 871081
Responsibility for the information and views set out in this report lies entirely with the authors.
The European Commission is not responsible for any use that may be made of the information it contains.

4. Appendix

N.A.

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 871081
Responsibility for the information and views set out in this report lies entirely with the authors.
The European Commission is not responsible for any use that may be made of the information it contains.

5. References

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 871081
Responsibility for the information and views set out in this report lies entirely with the authors.
The European Commission is not responsible for any use that may be made of the information it contains.