Project Title:	AQUACOSM: Network of Leading European AQUAtic MesoCOSM Facilities Connecting Mountains to Oceans from the Arctic to the Mediterranean
Project number:	731065
Project Acronym:	AQUACOSM
Proposal full title:	Network of Leading European AQUAtic MesoCOSM Facilities Connecting Mountains to Oceans from the Arctic to the Mediterranean
Type:	Research and innovation actions
Work program topics addressed:	H2020-INFRAIA-2016-2017: Integrating and opening research infrastructures of European interest

Deliverable No 4.3: Guidelines for database management, including controlled vocabulary

Due date of deliverable:	31 Mar 2018
Actual submission date:	28 Mar 2018, revised 08 April 2019
Version:	Version 2.0
Main Authors:	Eti E. Levi (AU), Süleyman Kazım Sömek (RF-SENS), Dennis Trolle (AU), Anders Nielsen (AU), Lisette de Senerpont Domis (NIOO/KNAW), Daphne Buijert-De Gelder (NIOO/KNAW), Simon Keeble (BLIT)

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 731065

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 731065

Project ref. number	**731065**
Project title	**AQUACOSM: NETWORK OF LEADING EUROPEAN AQUATIC MESOCOSM FACILITIES** **CONNECTING MOUNTAINS TO OCEANS FROM THE ARCTIC TO THE MEDITERRANEAN**

Deliverable title	Guidelines for database management, including controlled vocabulary
Deliverable number	D4.3
Deliverable version	Version 2.0
Contractual date of delivery	31 Mar 2018
Actual date of delivery	28 Mar 2018, revised 08 April 2019
Document status	Final
Document version	Version 2.0
Online access	Yes
Diffusion	Public
Nature of deliverable	Report
Workpackage	WP4
Partner responsible	AU/NIOO
Author(s)	Eti E. Levi (AU), Süleyman Kazım Sömek (RF-SENS), Dennis Trolle (AU), Anders Nielsen (AU), Lisette de Senerpont Domis (NIOO), Daphne Buijert-De Gelder (NIOO), Simon Keeble (BLIT)
Editor	Jens C. Nejstgaard (FVB-IGB), Project Coordinator
Approved by	Jens C. Nejstgaard (FVB-IGB)
EC Project Officer	Agnès Robin

Abstract	Through a Transnational Access program, the AQUACOSM project will collect data from aquatic mesocosms based on experimental facilities throughout Europe. As part of the AQUACOSM consortium agreement, partners have agreed to make data available, and this will be made searchable through a metadatabase. To ensure long-term data availability from these experiments, and to foster collaboration between researchers, this Guideline document describes best practices for curating data.

Keywords	Database management, metadata, data discovery, data availability, data storage, data formatting

This project has received funding from the *European Union's Horizon 2020 research and innovation programme* under grant agreement No 731065

Table of Contents

1. Executive summary

Mesocosm experimental data is diverse and standard practices are required to ensure that, as a community and Research Infrastructure, data is managed and made available for discovery and re-use in a consistent manner and in line with best practices for data / database management and metadata creation and maintenance. The purpose of this document is to outline the baseline requirements for data discovery, making data open-access, data storage, formatting and version control. This will form the basis upon which partners will be expected to comply with to ensure a consistent approach within the network of mesocosm infrastructures.

This document is an annex to the AQUACOSM Deliverable 4.2 [1], which defines the projects' Database Management Plan (DMP) in accordance with the H2020 Open Research Data Pilot. Based on Horizon2020 requirements, Deliverable 4.2 states that produced data should be findable, accessible, interoperable and reusable (FAIR). In this regard, AQUACOSM project briefly requires: 1) data (quality controlled and assured) to be open-access (no additional software, no password protection) within 6 months after publishable dataset is prepared, 2) both raw data and processed data (QA / QC) to be referenced with a Digital Object Identifier (DOI), 3) to follow AQUACOSM SOPs and guidelines for file naming conventions and version numbering, 4) to store the data in a trustworthy repository, and 5) metadata to be made available through the AQUACOSM website (http://aquacosm.eu/) on the metadata portal (see [1] for more details).

One of the most important provisions of the DMP is that project partners should follow the Ethics on research integrity (intellectual honesty and personal responsibility) as described in the Description of Action, GDPR and other such regulation. National or international legislation (e.g. on exotic or invasive species) on data collection should also be followed.

Co-funded by the European Union

2. Definitions and Terms

Archive: a place for the non-current data to be stored, a type of repository.

Backup: making a copy of any digital asset, to be sure to have a replacement in case of a hardware or software failure.

Controlled vocabulary: A collection of standardised or agreed upon terms that are used to identify parameters consistently.

Data: raw material that is used to create information.

Database: an organised data collection, mostly in digital form.

Database recovery: process to restore a database in case of a problem, such as loss or corruption.

Database security: protecting the data from unwanted access.

DMP: Database Management Plan.

EML: Ecological Metadata Language.

GitHub: a web-based version control system for both open source and private software repositories, originally designed for software source codes. It is commonly used for managing and sharing changes of big software projects, but also for other types of text based assets.

Metadata: data (description) that provides information about data.

Metadata standard: is developed for providing a common understanding of semantic components of a dataset and for accurately using and interpreting the data.

Metadata schema: the entire, defined, array of available metadata fields.

Quality assurance / quality control (QA/QC): processes used for error prevention in the dataset during and after the collection of data, thus preserving data quality.

Repository: a destination for storing metadata or data.

SOP: Standard Operating Procedure.

TA: Transnational Access.

XML Schema: is used to define the structure of data elements and relationships in an XML document.

XML: Extensible Markup Language, text -based scripting language, which determines the rules for depicting data structures.

3. Metadata & Data Discovery

3.1 Explanation of Metadata

Metadata is the information used to comprehend a dataset, simply put, “data about data”. A typical metadata describes “who, what, when, where, why and how (origin, organisation and features)” of an actual data set. Hence, users can discover, understand and use the actual data [2]. Metadata can be found everywhere. A library catalogue, with information on books and journals, for example, is a metadata.

Descriptive, structural and administrative metadata are the three main metadata types (Table 1);

- *Descriptive*: Metadata that is used for discovering and identifying information, like title and date, (e.g. Dublin Core).
- *Structural*: Metadata (e.g. XML) that is used to navigate and present information. It provides knowledge of how the components of data are organised and related, as it is in table of contents and indexes.
- *Administrative*: Metadata that is used for short- and long-term data management. It includes information on technical (e.g. on creation – when and how –, and quality control), rights management (e.g. access control) and preservation (e.g. archive) issues.

If there is a lack of detailed records of information on data (e.g. data collection date, location) then the actual data sets tends to fade (be forgotten) with time, especially after the original analysis of the actual data is completed. There is also a risk of losing the data and documentation (metadata) due to accidents, like crashed disks or fires, also to the leaving of a personnel who was in charge of managing the data [2]. Moreover incoherency of metadata or data records would complicate interoperability in scientific studies, if created metadata includes an array of different information, potentially resulting in non-standardised metadata presentation across studies, and in ecological data stored in many different formats [3, 4]. Therefore it is essential for researchers to have a detailed and standardised metadata document, and also to preserve/store both metadata and data in reliable locations.

Taking these possible risks into considerations, within the scope of the AQUACOSM project, a centralized metadata repository will be built following current standards in metadata vocabulary [1][4].

Metadata Type		Example Properties	Example
Descriptive metadata		Title Author Subject Publication date	World map ESRI ArcGlobe Satellite imagery map 22 Mar 2018
Structural metadata		Sequence Place in hierarchy	**Researcher**: Author, Institution → **Copyrights**: Institution, Year
Administrative metadata	Technical	File type File size Creation date/time	TIFF 70 KB 06 Feb 2018
	Preservation	Checksum[1]	1945687947
	Rights	Copyright status License terms Rights holder	©2012 ESRI http://www.esri.com/legal/software-license ESRI GIS company

[1] *Checksum*: it is a distinct identifier for the data and changes when data is modified.

Table 1. An example of metadata elements for descriptive, structural and administrative metadata (modified from [5]).

3.2 Benefits of high quality metadata for data discovery and re-use

The evident benefits of using and storing a high quality, standardised and structured metadata in a permanent archive, are:

- Preventing loss of information, thus ensuring long-term data availability
- Ease of data discovery and acquisition, since metadata is concise and can be used as a proxy for internet search
- Ease of understanding and interpretation of the data
- Simplified data sharing between researchers [4, 6].

3.3 Metadata platform – GeoNetwork

GeoNetwork OpenSource is a web-based mapping platform that is used as a cataloguing service for spatially referenced data, and for generating and managing metadata [7, 8]. The main objectives of this software are to reduce duplications, standardise information and data coherency, and improve data quality [7]. This platform also promotes collaboration (information sharing) among users by allowing users to access geospatial metadatabases and to search through the metadata generated for cartographic products.

Some of the main features include:

- Searching through the geospatial catalogues,
- Uploading and downloading any kind of content, such as data , documents,
- Interactive Web map viewer, combining Web Map Services from distributed servers,
- Online map layout generation and export in PDF format,
- Online editing of metadata,
- Scheduled harvesting and synchronisation of metadata between distributed catalogues (retrieved from [7]).

In addition, GeoNetwork supports various metadata standards used for geographic data (ISO19115/119/110) and for open data portals (Dublin Core) [9]. It, also, has a schema-plugin for implementing Ecological Metadata Language (EML) standard and since it has its own API, it can be harvested for interacting with other systems [9, 10].

Following a review of various platforms, or the inherent costs and risks of developing a bespoke solution, it was decided that GeoNetwork would be the most appropriate solution for the consortium and it is well established as a stable and compliant platform for metadata. Therefore, the AQUACOSM metadata platform will be based on GeoNetwork, as this will provide the AQUACOSM community with an already well established and rich metadata platform, ensuring sustainability beyond the lifetime of the project.

3.4 Common vocabularies

A controlled/common vocabulary include an organised, standardised and authorised list of terms [11]that have been agreed upon by a community of users [12]. In other words, it is based on rules guiding users on how to present their data. These vocabularies are specifically designed according to the subject of interest (e.g. oceanography). Glossaries, defining the natural-language terms, and thesauri, depicting the relationships of the terms (e.g. broader-term and narrower-term – see NERC website for an example) are the extensions of the vocabularies [11, 13].

If a project database uses natural language (i.e. speaking language), each user can enter the information as they want, however this may obstruct discovering some of the information by other users. For instance, consider that in a database both “Total Phosphorus” and “TP” are used, if users search only for one of these terms there is a good chance that they will miss some documents, which comprise the other term. Overall, common vocabularies are used to solve the difficulties related to ambiguous terms (e.g. synonyms) and the complexities arising due to the usage of natural language (e.g. misspellings) [11]. For this reason, AQUACOSM seek to employ a standard keyword vocabulary (controlled vocabulary) (see also [1]) and tailor it to mesocosm studies if needed [1].

3.5 Metadata standards

Every discipline has differing requirements for defining the metadata. Metadata standards have been developed so that vocabularies have standardised and clearly defined terms. The term 'location' is an example to the ambiguity in terms, since for one field of study only the name of the place may be enough, while for another coordinates may be needed [2]. These standards are important to ensure that users understand and interpret the data correctly, thus they facilitate data sharing and collaboration among researchers.

Many standards have been developed, such as for environmental sciences, education and finance. Some of the examples for the ecological and environmental standards are Dublin core Metadata Initiative (DCMI), ISO19115 and EML. DCMI is an open organisation and serves to a wide spectrum of communities (e.g. for business purposes) and ISO19115 is used for describing geospatial information [2, 14]. In addition to core standards, the AQUACOSM consortium will make use of the Ecological metadata language (EML), which is developed by the ecology discipline, as a metadata standard [1].

3.5.1 EML – use and limitations

Ecological Metadata Language was developed especially for ecological and environmental data [15], by the National Center for Ecological Analysis and Synthesis (NCEAS) and the Long Term Ecological Research Network (LTER) [3]. EML provides an extensible and flexible common structure (metadata standard) for the documentation of ecological data and use this structure for developing software applications [15]. It also enables automation of searching and retrieving information [16].

EML describes the most important aspects of ecological data, such as names, variable definitions, measurement units, identity of the responsible people for collecting the data, also geospatial (e.g. location of the research project and sample collection sites) and temporal (e.g. when the sampling was conducted) information. If needed it is also possible to describe the taxonomic information on the species as well. Information on the data usage restrictions (i.e. who can use it, when it can be used) is also provided. EML also provides a section for the data in tabular format, physical information, like file name, number of records and data table structure (i.e. column and row names) can be described [3].

Creating EML for an overall dataset (information on ownership, contact, spatiotemporal pattern and keywords) is relatively easy and fast (ca. 30 minutes), once users are familiarised with the basics of the ecological metadata. More detailed descriptions, for instance on the variables or on their definitions, can take longer time, depending on the structure of the datasets (e.g. interactions, number of variables). However, it is well worth for this effort, since detailed metadata means longer use and better availability of datasets for future research [3]. Therefore, the use of EML in AQUACOSM will enable users to easily search and find needed data, whilst at the same time facilitating the collaborations between researchers and providing a first step for future research opportunities.

It is noted that EML has two deficiencies: 1) It is a vast schema that attempts provide an all encompassing approach to cataloguing ecological metadata. The consortium are not expected to need all of the available functionality from the schema as it would be too 'heavy' in many cases. 2) It is potentially incomplete and is maintained as a 'community project'. Therefore, there could be deficiencies in capability when applied to the variety of mesocosm data within the consortium. We expect that it may be necessary to revise the standard and contribute to the 'community project'.

3.6 Distribution of metadata between networks

The storage of metadata in standardised format within GeoNetwork will allow other metadata portals to harvest the data and increase exposure and discoverability across networks outside of the AQUACOSM consortium. It also enables the metadata portal implemented as part of the AQUACOSM project to harvest metadata of relevance from external metadata services and enable a more extensive data discovery service.

3.7 Responsibilities for metadata maintenance

Metadata make a 'living document' and it is essential to ensure that it is created on time and the repository has up-to-date information. In order to keep the information up-to-date, versions, updates, deletions and insertions should be carefully conducted and recorded (history of the change). Moreover, using the same metadata repository will ensure that the related datasets are linked together. Upon completion of experiments carried out under AQUACOSM, metadata should be entered in the metadatabase as part of the Transnational Access (TA) requirements [1]. Metadata will be a priority output from all TA projects.

Besides responsibilities on maintaining metadata quality, project partners should also be sure on the quality of both metadata and data, backups should be taken against any type of problems and most importantly metadata should be stored in the AQUACOSM centralized repository and data in an open-access decentralized repository, according to the requirements of DoA / Consortium Agreement (Article 29.3 in [17]. Responsibilities of the project partners can be summarised as in Fig. 1.

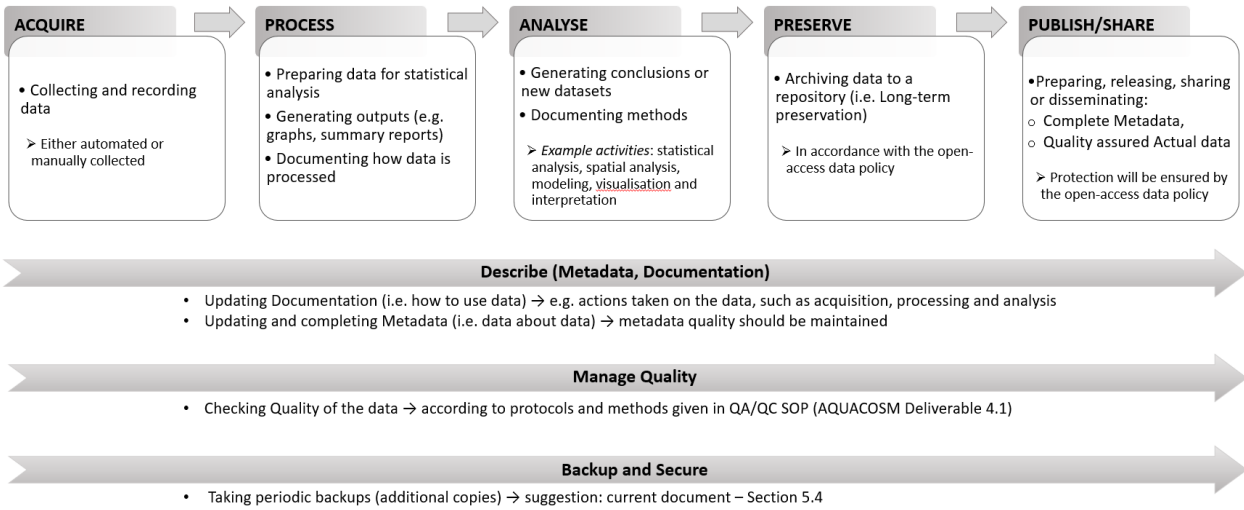

Figure 1. Responsibilities of the project partners for ensuring a solid DMP (modified from [17])

4. Data availability and open-access

4.1 Open-access data

Horizon2020 Annotated Model Grant Agreement defines Open-Access data as having "the right to access and reuse digital research data under the terms and conditions set out in the Agreement" [19]. Simply put, Open-Access means data must be publicly available, but only with taking necessary precautions. H2020 encourages Open-Access data, because collaboration between scientists and future work by improving previous studies underlie modern research.

Open-Access helps by:

- Increasing result quality, because it allows building-up earlier research outcomes,
- Promoting cooperation between researchers, so reduces the effort spent,
- Speeding up innovation,
- Encouraging participation of citizens and society to scientific research [19].

4.2 Requirements for making data available publicly

Making the data Open-Access within the AQUACOSM community is one of the aims of DoA / Consortium Agreement (Article 29.3 in [17]). **This agreement states that recipients of the funding must store their data, associated metadata and information on the data collection (e.g. tools, instruments) in a repository, so that it can be accessed, used and disseminated free of charge**. Also, metadata and data sharing should be carried out as soon as possible, within the projects' data management plan deadlines. The AQUACOSM partners will follow this agreement and as part of Transnational Access (TA) it is a requirement that data should be made freely available after completing the experiment and within 6 months a publishable dataset is prepared [17], in general, see however 4.3.

4.3 Exceptions for making data publicly available

Only based on specific criteria, there can be exceptions to making data publicly available. AQUACOSM partners have the option to choose to partially or entirely renounce from sharing the data if Open-Access will jeopardise the achievement of their research or the project. It is important for project partners to consider their own legitimate interests to adequately protect the results within the conditions presented in the Consortium Agreement and it is also possible to put an embargo for a period of time for publishing the data, but partners who are willing to put an embargo period for publishing their data should justify the reasons for this [17]. When and how to publish the results or the reasons of exceptions for publishing them, differs for each project. To exemplify, we will allow for an embargo of three years after finalization of the AQUACOSM TA to allow for completion of a PhD thesis [1].

4.3.1 Examples

Exceptions or restrictions may be put in place where a legitimate reason exists, such as the protection of commercial or industrial exploitable results that may create unnecessary competition if released too early or the embargo required for a PhD student to complete a thesis.

4.3.2 Procedure for exceptions (SC authorisation)

To allow for consideration of exemption, AQUACOSM partners / TA participants who are unable to make their data available for any reason are required to submit an official request to the Steering Committee and the Committee shall decide if the exemption will be implemented or rejected. In case there are any uncertainties according to valid EU-regulations the Steering Committee may seek advice from the EC Project Officer before the matter is decided.

5. Data storage and Archiving

5.1 Distributed data networks and storage

Data repositories are platforms that are used to store actual data [4]. Centralized and decentralized infrastructures are the most common approaches for designing the repositories (i.e. how the data is interconnected). In a centralized design the technology and support is handled by a single system. Hence, every service (e.g. data search, finding, security and management) provided by the repository are controlled from the centre [20]. Distributed data networks do not have a central repository and a decentralized infrastructure form their basis. In this decentralized, or distributed, design, more than one system conducts the management of the technology and support. This also means that each system has its own repositories to store data [20].

Providing interconnection between metadata-based databases with actual data is important, because it will allow wider-scale usage of data [4]. In other words, interfaces between various data management tools and repositories enables researchers to achieve maximum efficiency in data sharing. Within the scope of the AQUACOSM project data should be openly accessible. Metadata sharing will be through a centralized repository embedded in the AQUACOSM website (http://aquacosm.eu/) and data sharing will be through decentralized (distributed) repositories that each institution will choose according to their own needs [1].

5.2 Short term and long term storage

Fundamental components of a well-grounded database management plan are “short-term data storage” and “long-term data preservation”. Storage place for data and associated code, and how to store this should be considered during a management plan [21].

Data storage is an important part of managing data and it refers to safe keeping of information in the memory (i.e. computer media) for later use. There are various storage media for short-term storage of the data during the course of a project, including computers, external hard drives, also drivers and servers provided by local institutions, though all comes with risks [22].

Long-term preservation is defined as permanent data archiving and it especially covers the period following completion of a project, when data is not actively used. Preservation requires keeping information, such as data and associated codes, available and without corruption [22]. The best choice for long-term data preservation is to store the information in a reliable center or repository that is suitable for the research area of interest [21, 22].

A good data storage strategy (i.e. where and how to store data) helps scientists to significantly reduce the risk of losing the data and long-term preservation enables them to search, discover and use the data, as well as facilitates collaborations among researchers. For long-term preservation it is important to find reliable repositories, like decentralized (distributed) repository Knowledge Network for Biocomplexity (KNB), which is a Member Node of the DataONE framework [16]. To be able to provide a good storage option metadata produced during the course of the AQUACOSM project will be made freely available through a centralized repository (http://aquacosm.eu/) and each partner is responsible to make their data openly available through a decentralized repository by their choice, though some restrictions for data sharing may apply if necessary (see details in [1]).

5.3 Storage formats and database technologies

Given the complexity of different database technologies and storage formats, it is outside the scope of this document to specify for partners 'how' data should be stored other than to stress the following:

- Data should be stored in a format that will enable ease of reuse without specific software tools.
- Plans should be in place for short-term (after experiment), long-term (online availability) and long-term archival of collected data.
- Data should be subject to change management, quality control and appropriate security measures put in place to ensure the integrity of the data.
- All data should be allocated DOIs or other digital identifiers to ensure correct citation of data.

5.4 Backup & Recovery

A robust and tested backup & recovery plan should be in place for all data. All data should be recoverable and available again within a **maximum** of 48 hours from loss of service. However, this should be considered the absolute maximum. A good backup and recovery plan should ensure that the data is recoverable within minutes from a recent failure with perhaps extended timescales from long term archive or offsite backups.

It is outside the scope of this document to prescribe the exact methodology and policies, as these are generally managed at institutional level, but it is important to note that each partner is responsible for ensuring such a robust strategy is in place.

6. Data Formatting

6.1 Data storage formats

One of the most important elements in database management is to preserve the data for long-term use, thus data should be stored in suitable formats. In order for computers to read the data files, even in the future, data formats should be standard, easy to understand and open. The suitable file type would change depending on the type of data to be stored (e.g. numeric, text or images) [23]. These features enable higher interoperability, easy usage and increased data longevity, also prevent errors that can be resulted during conversions between formats [23, 24]. These type of data formats should not require a specialised software to open files. The ideal option is not to use proprietary softwares and hardwares, or not to purchase a commercial license, because proprietary formats change or organisation that maintain these formats may go out of business, making these formats too expensive to be afforded or risky for users [23]. Therefore, problems that can occur due to a change in software version (e.g. new version-not possible to open old documents) will not occur when open data formats are used. Some of the recommended open and machine readable formats for preserving ecological data are, for text data types '.txt', for tabular types '.csv', for geospatial types 'Shapefile' and 'GeoTIFF', and for documents 'Extensible Markup Language (XML)' and 'JavaScript Object Notation (JSON)', scanned images on the other hand are not easy to process.

6.2 Data transfer formats

Transferring data means to move or copy a dataset from one place to another, for example between networks, people or between applications. Compatibility is a very important feature for this process, because data should be read easily by the receiving side. Therefore data transfer formats should be standardised. Extensible Markup Language (XML), JavaScript Object Notion (JSON), NetCDF, ODV, CSV etc. are some examples for the formats that are widely used for transferring the data.

7. Quality Assurance / Quality Control

Quality Assurance and Quality Control (QA&QC) in data are processes for ensuring and maintaining high data quality by preventing errors from entering a data set before and after data collection [25]. Quality assurance aims at defining standards before starting to collect data. As an example in AQUACOSM, standardization will include "parameter name, parameter formats, measurement units, codes and metadata". Moreover, a person who will be responsible for data quality should be assigned [26]. After producing the data, quality control should be conducted to check if there is any inconsistencies in the dataset. Raw data (original data set) must be saved before any changes are made in the data. QC include checking the range of the values, any gap or missing values, irrational results (i.e. any that are very different than expected) and outliers (i.e. outside of the expected pattern) (see [26] for details). QC should also be conducted on the sensor network, including checking date and time, data range, persistence (if same values recorded), change in slope, and also internal and spatial consistencies (see [26] for details).

8. Version Control

8.1 Data version control

Maintaining the version of data collected is important to ensure that future work conducted on the data are carried out on the correct version of the data.

Data collected from a sensor (raw data) and data made available for analysis, for example, will often be different versions of the same data. It is important to ensure that data is versioned and metadata created that relates to each version of the data. If data is subsequently updated, following identification of a systematic error in retrieval for example, it is important that the versions and metadata records are updated accordingly.

8.2 Digital Object Identifiers (DOI)

A Digital Object Identifier (DOI) refers to "digital identification of an object" and it is a unique and permanent identifier. DOI can be used to distinguish contents in a digital environment, like the internet. It is an alphanumeric string and it provides a persistent link to the current location of a specified object on the internet [27]. DOIs are generally used together with version control mechanisms for ensuring that the data will not disappear [28].

Assigning a DOI to a dataset is one of the steps to solve the problems resulting from the change of data locations or ownership. Moreover, data can be cited by the users through DOIs and this increases the discoverability of the data for further research. Data producers also benefits from DOIs, because when their data is cited, their research receives a proper recognition [29].

All of the data, both the raw and processed (quality checked), should have a DOI within the scope of AQUACOSM project [1].

9. References

[1] de Senerpont Domis, L.N., Başoğlu, D. Beklioğlu M., Culina, A., Keeble S., Vidussi-Mostajir, F., *et al.*, (2018). *AQUACOSM D4.2, Database Management Plan adhering to the H2020 Open Research Data Pilot*. https://www.aquacosm.eu/project-information/deliverables/. Accessed 20 Mar 2018.

[2] Michener, W.K., (2018). Creating and Managing Metadata. In Recknagel, F. and Michener, W.K. (Ed.), *Ecological Informatics Data Management and Knowledge Discovery*. Springer International Publishing AG, Switzerland, pp. 71-89.

[3] Fegraus, E.H., Andelman, S., Jones, M.B. and Schildhauer, M., (2005). Maximizing the Value of Ecological Data with Structured Metadata: An Introduction to Ecological Metadata Language (EML) and Principles for Metadata Creation. *Bulletin of the Ecological Society of America*, 86:158-168.

[4] Jones, M.B., Schildhauer, M.P., Reichman, O.J., and Bowers, S., (2006). The new bioinformatics: integrating ecological data from the gene to the biosphere. *Annual Review of Ecology, Evolution, and Systematics*, 37:519-544.

[5] NISO, (2017). Understanding Metadata. National Information Standards Organization, Baltimore

[6] Gordon, S. and Habermann, T., (2018). The influence of community recommendations on metadata completeness. *Ecological Informatics*, 43:38-51.

[7] GeoNetwork Opensource, (2017). GeoNetwork User Manual (Release 2.10.4-0). https://geonetwork-opensource.org/manuals/2.10.4/eng/users/GeoNetworkUserManual.pdf. Accessed 22 Mar 2018.

[8] DataONE, (2018). data observation network for earth. https://dataone.org. Accessed 22 Mar 2018.

[9] GeoNetwork Opensource (2018). https://geonetwork-opensource.org/. Accessed 22 Mar 2018.

[10] Oggioni, A., Wohner, C., Watkins, J., Ciar, D., Schentz, H., Lanucara, S., *et al.*, (2017). D3.1 eLTER State of the art and requirements. European Long-Term Ecosystem and SocioEcological Research Infrastructure.

[11] Madin, J.S., Bowers, S., Schildhauer, M.P. and Jones, M.B., (2008). Advancing ecological research with ontologies. *Trends in ecology & evolution*, 23(3):159-168.

[12] Parr, C.S. and Thessen, A.E., (2018). Biodiversity Informatics. In Recknagel, F. and Michener, W.K. (Ed.), *Ecological Informatics Data Management and Knowledge Discovery*. Springer International Publishing AG, Switzerland, pp. 375-401.

[13] NERC, (2018). NERC Vocabulary Server. https://www.bodc.ac.uk/resources/products/web_services/vocab/. Accessed 22 Mar 2018.

[14] ISO, (2018). ISO 19115-1:2014. Geographic information -- Metadata -- Part 1: Fundamentals. https://www.iso.org/standard/53798.html.

[15] Blankman, D. and McGann, J., (2003). Ecological metadata language: Practical application for scientists. http://im.lternet.edu/sites/im.lternet.edu/files/emlHandbook.pdf. Accessed 22 Mar 2018.

[16] KNB, (2018). The Knowledge Network for Biocomplexity. https://knb.ecoinformatics.org/#. Accessed 22 Mar 2018.

[17] European Commission, (2017). Annotated Model Grant Agreement: V4.1 – 26 October 2017 http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/amga/h2020-amga_en.pdf. Accessed 22 Mar 2018.

[18] USGS, (2018). USGS Data Management. https://www2.usgs.gov/datamanagement/why.php. Accessed 20 Mar 2018.

[19] European Commission, (2017). Guidelines to the Rules on Open Access to Scientific Publications and Open Access to Research Data in Horizon 2020. V3.2 – 21 March 2017. http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf. Accessed 22 Mar 2018.

[20] Kowalczyk S. and Shankar K., (2011). Data sharing in the sciences. *Annual Review of Information Science and Technology*, 45: 247–294.

[21] Michener W. K., (2018b). Project Data Management Planning. In Recknagel, F. and Michener, W.K. (Ed.), *Ecological Informatics Data Management and Knowledge Discovery*. Springer International Publishing AG, Switzerland, pp. 13-27.

[22] Briney, K., (2015). Data Management for Researchers: organize, maintain and share your data for research success. Pelagic Publishing Ltd., Exeter, UK.

[23] Hart, E.M., Barmby, P., LeBauer, D., Michonneau, F., Mount, S., Mulrooney, P., *et al.*, (2016) Ten simple rules for digital data storage. *PLoS Computational Biology*, 12(10): e1005097.

[24] Cook, R.B., Yaxing, W. Hook, L.A., Vannan, S.K.S. and McNelis, J.J., (2018). In Recknagel, F. and Michener, W.K. (Ed.), *Ecological Informatics Data Management and Knowledge Discovery*. Springer International Publishing AG, Switzerland, pp. 89-115.

[25] Michener, W.K. and Jones, M.B., (2012). Ecoinformatics: supporting ecology as a data-intensive science. *Trends in Ecology and Evolution*, 27:85-93.

[26] AQUACOSM Deliverable 4.1 Standard Operating Protocol for QA-QC.

[27] Paskin, N., (2010). Digital object identifier (DOI) system. In *Encyclopedia of Library and Information Sciences*, 3rd ed. Taylor & Francis.

[28] Porter, J.H., (2018). Scientific Databases for Environmental Research. In Recknagel, F. and Michener, W.K. (Ed.), *Ecological Informatics Data Management and Knowledge Discovery*. Springer International Publishing AG, Switzerland, pp. 89-115.

[29] Michener, W.K., (2018c). Data Discovery. In Recknagel, F. and Michener, W.K. (Ed.), *Ecological Informatics Data Management and Knowledge Discovery*. Springer International Publishing AG, Switzerland, pp. 115-129.