Project	MONOCLE H2020 (grant 776480)	Start / Duration	1 February 2018/ 48 Months
Dissemination	PUBLIC	Nature	ORDP
Date	20/11/ 2018	Version	1.1

MONOCLE

Multiscale Observation Networks for Optical monitoring of Coastal waters, Lakes and Estuaries

Deliverable 5.1

Data management plan (v1.0, July 2018)

Project Description

Funded by EU H2020 MONOCLE **creates sustainable *in situ* observation solutions for Earth Observation (EO) of optical water quality in inland and transitional waters.** MONOCLE develops essential research and technology to lower the cost of acquisition, maintenance, and regular deployment of *in situ* sensors related to optical water quality. The MONOCLE sensor system includes handheld devices, smartphone applications, and piloted and autonomous drones, as well as automated observation systems for e.g. buoys and shipborne operation. The sensors are networked to establish interactive links between operational Earth Observation (EO) and essential environmental monitoring in inland and transitional water bodies, which are particularly vulnerable to environmental change.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776480

Deliverable Contributors:	Name	Organisation	Role / Title
Deliverable Leader	Oliver Clements	PML	WP5 Developer
Contributing Author(s)	Kathrin Poser	WI	WP5 Lead
	Stefan Simis	PML	Contributor
Reviewer(s)	Kathrin Poser	WI	WP5 Lead
Final review and approval	Stefan Simis	PML	

Document History:

Release	Date	Reason for Change	Status	Distribution
0.1	18/7/2018	Initial document	Draft	Internal
0.2	20/7/2018	Data summaries and purpose added per data type	Draft	Internal
0.9	30/7/2018	Complete draft for review	Draft	Internal
1.0	01/08/2018	Reviewed document for release	Completed	EC
1.1	20/11/2018	Digital Object Identifier added	Released	Public

To cite this document:

Clements, O., Poser, K., Simis, S.G.H. (2018). D5.1 Data Management Plan. *Deliverable report of project H2020 MONOCLE (grant 776480)*, doi: 10.5281/zenodo.1492265

© The authors. Published under a Creative Commons Attribution-NonCommercial-NoDerivatives license.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776480

TABLE OF CONTENTS

Acronyms 4
1. Executive Summary 5
2. Scope 5
3. Introduction 5
4. Data summary 6
Data purpose and utility 6
Data Types 7
Citizen generated data 7
Automated data collection 8
Manual data collection 8
Earth observation data 8
Image data collected with Remotely Piloted Aircraft Systems (RPAS) 9
Pre-existing data 9
Research results and derived data sets 9
5 FAIR data principles 9
General data documentation and guidance 10
Metadata 10
Data Access, Interoperability and respecting Intellectual Property 10
Data Sharing and Reuse 11
Data Preservation and Archiving 11
Data Register 12
6 Allocation of resources 13
7 Data security 13
8 Ethical aspects 13
Appendix 14
Data Register Template 14

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 776480

Acronyms

DMP	Data Management Plan
EO	Earth Observation
ESA	European Space Agency
FAIR	Findable Accessible Interoperable Re-usable
GIS	Geographic Information System
GEOSS	Global Earth Observation System of Systems
ISO	International Organisation for Standardisation
NASA	National Aeronautics and Space Administration
OGC	Open Geospatial Consortium
OLCI	Ocean and Land Colour Imager
PML	Plymouth Marine Laboratory
RPAS	Remote Piloted Aircraft System
SOS	Sensor Observation System
UAV	Unmanned Aerial Vehicle
UTC	Coordinated Universal Time
WCS	Web Coverage Service
WFS	Web Feature Service

1. Executive Summary

The objective of this data management plan (DMP) is to detail the plan for management of data generated and collected within the MONOCLE project. The DMP describes the data management life cycle for all datasets collected, processed and/or generated by the project. It covers

- what data will be collected, processed or generated
- how the data will be handled during and after the project
- who is considered as an owner of a data set and who it is shared with
- how the sharing of the data within and outside the project is organised
- what formats, meta data and standards the data will adhere to
- how data will be curated and preserved

MONOCLE data will consist of a diverse range of types and formats and standardization of the data and data flows is one of the main project objectives.

2. Scope

This document, The Data Management Plan, is intended for internal and external use, describing the mechanisms that MONOCLE will put in place to ensure all public data follow the FAIR (Findable, Accessible, Interoperable, Re-usable) data management principles. This is a living document, updated periodically to reflect new data sets made available through MONOCLE. The current document presents the status and planning at month six of the four-year project.

Streamlining data access and interoperability from the sensor to the user is one of the main aspects of the MONOCLE project. Therefore, a number of related reports which detail the methodologies developed in the project will be of interest. At present, D5.2 outlines the data infrastructure and standards that will be implemented in the project. D5.4 will describe the final implementation of the data flow between MONOCLE subsystems.

3. Introduction

The aim of MONOCLE is to implement enabling technologies for the deployment, management and maintenance of sensors and sensor networks. Sound data management is pivotal for fully realising the benefits of MONOCLE. Well curated data will stimulate and ensure smooth collaboration between the project partners and will allow the users to easily evaluate and put to use the data received from the project. For dissemination and exploitation, open access to data generated in the project will help to underpin the credibility and stimulate uptake of MONOCLE results.

The MONOCLE project will follow the FAIR (Findable, Accessible, Interoperable, Re-usable) data paradigm and this is reflected in the data management plan. Data will be **F**indable through the various user applications that interface to the data services provided by the MONOCLE back-end. A web based geographic information system (GIS) will be publicly available with a data search feature acting on parameter, spatial and temporal coverage or data originator fields. Appropriate datasets

will also be registered in public archives such as ZENODO and GEOSS, this will enhance their ability to be found and **R**e-used, even if the MONOCLE back-end should cease to operate. The use of data services designed for system inter-operability will guarantee that all open data within the project are widely **A**ccessible now and in future. The user applications (web based and in the form of source code) will also improve **A**ccessibility with focused information available, including through intuitive tools. **I**nteroperability is also made possible through the use of common data formats and standardised data services. For instance, it would not matter what format the original data are stored in when requested via a Sensor Observation Service as the response is documented and standardised.

The DMP will be updated as a "live" document during the lifetime of the project, with four scheduled release dates. Document D5.2 "System architecture and standards report", accompanies the first release of the DMP and describes data sources and interfaces in additional detail.

4. Data summary

Data purpose and utility

Observation of global coastal and inland water bodies with ocean-colour satellite sensors has reached full operational potential through the latest satellite missions in the Copernicus programme. The global societal demand for water quality information through downstream EO services is increasing and expanding into domains of public health, agriculture, aquaculture, energy and food safety, drinking water, conservation of ecosystems and biodiversity conservation, navigation and recreational use of water resources. Inland and transitional water bodies, however, represent a staggering range of optical and environmental diversity. A dedicated concept for EO-supporting in situ services for optically complex waters is necessitated by the limited ability of present in situ activities to add additional value to operational EO missions.

To improve in situ components of the GEOSS and Copernicus services in optically complex waters, MONOCLE will introduce new sensor technological development across a range of innovative platforms. MONOCLE will combine high-end reference sensors in a spatially sparse configuration with a complementary, higher density, network of low cost sensors for smartphones and unmanned aerial vehicles (UAV or drones).

The full MONOCLE sensor suite and the data gathered and processed with MONOCLE sensors and processing means will serve the EO research communities for water and atmosphere with a rapidly replenishing volume of reference observation data, reducing both local-regional (improved atmospheric correction) and global (improved algorithms) observation uncertainty. The MONOCLE integrated observation service concept, particularly when integrated with EO services, significantly lowers the technology and computing requirement for innovators in environmental observation in general, and water quality management in particular. This reduction is critical for uptake and engagement in developing regions. By making both data and supporting software openly available the project will boost innovation with app developers, environmental consultants, data analysts and visualisation artists worldwide.

The open data strategy of MONOCLE plays a central role in opening opportunities to the EO sector, not merely in Europe but also in supporting downstream users and regional information providers in data-poor regions, particularly developing countries. For the latter, MONOCLE will lower the threshold for computational and technological capacity to actively contribute to the global observation system

Data Types

MONOCLE will collect a wealth of data on water quality from multiple sources that can be categorised as one of:

- In-situ data, either:
  - Data collected by non-expert participants (e.g. citizen scientists)
  - Data from automated instruments (e.g. on buoys, ships)
  - Data from manually operated instruments (e.g. hand-held sensors, piloted aircraft)
- Satellite data of inland, transitional and coastal water bodies
- Image data collected with Remotely Piloted Aircraft Systems (RPAS)
- Pre-existing data, accessed in (external) databases or directly contributed by stakeholders
- Research results and derived data sets

Each of these data sources has specific characteristics and challenges which are summarised below:

Citizen generated data

MONOCLE will engage with groups of volunteers in citizen science campaigns where the citizen scientists collect water quality data and submit these to the MONOCLE system via a mobile app. A number of different parameters will be collected by the citizen scientists, either ad-hoc or during larger campaigns. Such campaigns can deliver a large amount of data in a fixed time period, but are more difficult to plan as motivation of the volunteers is pivotal. A fundamental principle of MONOCLE data management is that the apps used to collect data will also have access to stored results, providing immediate feedback where possible.

Citizen participation requires additional ethical considerations, which are discussed further below.

Citizen observations are presently foreseen to be collected through the Earthwatch Freshwater Watch app and the iSPEX app (under development). Data exchange formats are still being decided upon, but will likely follow concepts of interoperability using Web Feature Services (WFS) and Sensor Observation Services (SOS), with the MONOCLE back-end communicating with the respective data stores of iSPEX and Earthwatch. Hence, no ‘raw’ data format is currently considered here.

The global Freshwater Watch dataset currently contains > 20,000 datasets, where each contributor is represented as a separate dataset. For iSPEX, the main mode of operation is foreseen to be in dedicated campaigns. Data volumes associated with the FreshWater Watch are modest as these take the form of forms and occasionally photos, likely to remain in the order of Gigabytes or less. The iSPEX collects a range of smartphone camera photos, likely to range in the order of Gigabytes. Data storage at this magnitude is not currently seen as an issue.

Automated data collection

A variety of automated sensors will be deployed by project partners such as radiometers, fluorometers and absorption meters. The sensors can be deployed at fixed positions (e.g. on buoys, poles or jetties) or on moving platforms (ships, RPAS). Data will in general be collected at high frequency and immediately transmitted to the MONOCLE system. However, if deployed in remote locations, the sensors can also collect data less frequently, and store measurements if they are not online. Data acquisition, processing and transmission should all be automated with these sensors as should be quality control mechanisms. The aim of MONOCLE is to provide these sensors with interactive interfaces so that measurements can be triggered, sensors turned on and off or calibrations performed remotely.

During and / or following data collection, most of the optical sensors require data processing, calibration and quality control. The intention is for these processes to be highly automated, with suspect data flagged as not recommended for use and to be inspected by the data creator / curator.

Where existing data stores are considered and an application programming interface (API) is not already in place, one will be created. The SOS interface will be preferred in the development of new communication interfaces from individual sensors. In addition, an SOS compliant data wrapper will be made available for legacy sensors.

Automated high-frequency data collection for the MONOCLE sensors is estimated in the order of tens of megabytes per observation day. Transfer, storage, and dissemination of these volumes is not currently seen as an operational issue.

Manual data collection

The project will collect data in field situations using handheld and manually operated instruments. This includes new sensors intended for short-term deployment, and high-end reference sensors operated only during validation campaigns, following described protocols. In all cases, the measurement records will be referenced by geo-location and UTC time-stamp and the measurement protocol which was used. These measurements are subject to further quality assurance (protocols) and quality control by the operators.

Manual data collection during field campaigns is estimated to deliver in the order of several gigabytes of data per campaign. Transfer, storage, and dissemination of these volumes is not currently seen as an operational issue.

Earth observation data

While the focus of MONOCLE is on providing a network of in situ observation to support Earth observation of optical water quality, for demonstration of the use and benefits of the MONOCLE services for Earth Observation services, satellite-derived data will be produced within the project. For dedicated case studies, high resolution (Sentinel-2 MSI) and medium resolution (Sentinel-3 OLCI) data will be acquired and processed into water quality information products making use of MONOCLE in situ data for calibration and validation.

Data storage needs for EO data for the selected MONOCLE regional use cases (Lake Balaton, Scottish Lochs, Danube Delta, Lake Tanganyika, and several smaller sites) is in the order of hundreds of gigabytes of data per year, which has been costed in the project budget.

Image data collected with Remotely Piloted Aircraft Systems (RPAS)

The purpose of data collection with RPAS is to construct mosaic maps of waterbodies from which water quality parameters can be derived. The RPAS systems may also serve as direct reference to satellite data, with the added advantage of detailing fine spatial features which can explain aberrations in processed satellite data, where fine features are not directly visible due to a large pixel size. The raw image data are too large (and not useful) to be disseminated beyond the data processing centres, where they are archived on suitable storage media (e.g. tape drives). Processed parameter-specific maps will be disseminated through the MONOCLE data back-end using machine interfaces (WCS).

Storage and archiving needs associated with the image data are in the order of terabytes of data and budgeted for in the project.

Pre-existing data

Pre-existing in situ datasets will consist of collections of optical and biogeochemical measurements contributed by various stakeholders, either as independent data sets where MONOCLE is given a licence to use and distribute these, or as part of curated data bases (e.g. LIMNADES for inland water). Access constraints will be recorded and maintained as part of the registration of the data set in the MONOCLE back end.

Pre-existing data will also take the form of large scale satellite data archives downloaded from space agencies (ESA and NASA), which are then used for further processing to a usable format, in turn integrated with data from MONOCLE sensors.

Research results and derived data sets

In the process of research and development, outputs will be generated in the form of publications, presentations, tables and datasets, survey results. Such results will be stored in the project management portal for access within the project consortium, the size not likely to exceed 100Mb per item. Public reports will be available also through the website. Public deliverables will be available through the website and OpenAire. The methodology is detailed in D9.3 “Open data repositories”.

The open access requirement for H2020 publications will be honoured through either the green or gold open access route. Each project partner is responsible for delivering publications through their chosen open access route – open access publication fees are an eligible project cost. In addition, these papers will be included in the Zenodo/OpenAire repository that has been set up for MONOCLE.

5 FAIR data principles

All MONOCLE research data will be curated according to the 'FAIR' principle, i.e. Findable, Accessible, Interoperable and Re-usable. In the following, a short overview is given of the building blocks to

reach this goal. It should be noted that due to the ongoing development of the system, further detail will be provided in future releases of this document. The following are guiding principles.

Details on each data set will be kept in a central data register, discussed further below.

General data documentation and guidance

Any documentation such as measurement protocols, system descriptions and use cases will be linked within the data register and a copy will be kept in the MONOCLE back-end, where possible. Users of the MONOCLE front-end will be able to access these documents when accessing a corresponding data set.

Where data sets are 'frozen' to create a snapshot of available data at a given point in time, these datasets will be versioned and uploaded to public repositories providing a digital object identifier.

By default, all data generated in MONOCLE will be openly available (see Data Access, below), with the exception of unprocessed, uncalibrated data if these have no value to the user. Such data will nevertheless be stored and curated. Data contributed from external sources are the exception to this rule. In such cases, data ownership and licensing will govern whether dissemination beyond MONOCLE is possible.

Metadata

Initially the metadata profile ISO 19115 will be used to describe datasets that are made available. As a common ontology the CF conventions (cfconventions.org) will be followed or extended. These metadata conventions ensure that data are identifiable, usually as part of (live) data streams, using appropriate search terms and key words.

Additional metadata requirements to enable MONOCLE data interoperability developments are described in D5.2 "System architecture and standards report". These requirements concern data ownership, licensing, access restrictions (embargo periods), as well as geospatial parameters. The definition of the minimum and recommended metadata for MONOCLE data sets will be refined during the implementation of MONOCLE WP5.

A guiding principle for MONOCLE sensors and platforms is that metadata are injected into the data flow at the point of measurement, either at the sensor or using a dedicated sensor interface.

Data Access, Interoperability and respecting Intellectual Property

Processed data intended for public access and not subject to ethics limitations will be made available through the Open Geospatial Consortium (OGC) Sensor Observation Service (SOS), Web Feature Service (WFS) and Web Coverage Service[1] (WCS) standards initially with other standard and bespoke data interfaces being added as required. MONOCLE aims to apply these standards to communication

1 : http://www.opengeospatial.org/standards/wcs

between sensors, sensors and data hubs, the MONOCLE data back-end, front-end and user applications. Details can be found in D5.2 “System architecture and standards report”.

Data generated as part of MONOCLE will be free of cost. Data access restrictions and intellectual property rights will, however, remain as set by the dataset creator/owner where applicable. Unless specified, all data will be treated as FAIR open data. In practise, the following data access levels are foreseen:

- open access, not requiring registration, providing access to data identified as open without license restrictions
- limited access, requiring registration, providing access to open data as well as data sets with a limited license for use (e.g. non-commercial, accrediting ownership, delayed release etc).
- restricted access, requiring registration, providing access to data owned by the user and any data sets this specific user has been granted access to.

Any software tools that are required to access and, to a limited extent, make use of the data, which are developed during MONOCLE, will be available free of cost through software repositories such as those already set up on Zenodo and Github (see D9.3 “Open data repositories”, for details). Essential software tools required to make use of the data have not been defined at this point.

Data Sharing and Reuse

All data accessible through the MONOCLE backend data services (Sensor Observation Service, Web Feature Service, Web Coverage Service) will be publicly findable, with accessibility rules based on ownership and licensing defined drawn from the metadata and data register. Any data producer that requires data to be delayed in its release will carry that information such that it can be securely stored and released only when appropriate. In many use cases, however, it will be beneficial to the user to know that embargoed data exist. Such data embargoes will feature in the extended metadata and the data register and can take the following shapes:

- restricted data that are identifiable by measurement parameter, and as collected within a given geographical range and time period
- restricted data identifiable as above but including exact time and location of observation
- restricted data identifiable as above, but including information about the data owner

Data Preservation and Archiving

Data will be kept available for a minimum of three years after the end of MONOCLE. Beyond this period, e.g. if the service should no longer be deemed useful or sustainable, data will be archived at a secure open access location, insofar as data licensing permits. The project intends to create links between MONOCLE data service and large-scale public data archives (e.g. GEOSS) for long-term accessibility.

Requests to remove a data set from the MONOCLE services can be submitted to the Coordinator and will be handled in a manner equivalent to the GDPR for personal data.

Within the project, Work Package 8 is dedicated to planning for long-term sustainability and evolution of MONOCLE from a service concept into an operational in situ service. Each of the development and innovation activities have set deliverables that will be conditioned by the identified end users and stakeholders through early stage trend and gap analysis. The input from the sensor manufacturing industry (beyond those well represented in the consortium), by primary in situ data producers (e.g. environment agencies), or primary data consumers (e.g. EO service developers) will provide both a cornerstone and vision for development. Commercial sensor and service development will be explored to support a 180 degree market perspective for MONOCLE system components and branding as a whole, exploring manufacturing chains and economies of scale, IP licensing, and patent searches, where applicable. Public-private partnerships and corporate sponsorships (providing green credentials) to sustain citizen observatories and management of 'super sites' will be considered in this work, delivered as an evolving exploitation plan.

Data Register

The data register will be maintained as a "live" document; a snapshot will be created for each DMP release. A template is included in the Appendix.

The data register is based upon information and restrictions supplied by the upstream data provider matched to Horizon 2020 guidelines as below (in *italics)*:

- ***Data set reference and name***
  Identifier for the data set to be produced.

- ***Data set description***
  Descriptions of the data that will be generated or collected, its origin (in case it is collected), nature and scale and to whom it could be useful, and whether it underpins a scientific publication. Information on the existence (or not) of similar data and the possibilities for integration and reuse.
  - *Standards and metadata*
    Reference to existing suitable standards of the discipline. If these do not exist, an outline on how and what metadata will be created.
  - *Data sharing*
    Description of how data will be shared, including access procedures, embargo periods (if any), outlines of technical mechanisms for dissemination and necessary software and other tools for enabling re-use, and definition of whether access will be widely open or restricted to specific groups. Identification of the repository where data will be stored, if already existing and identified, indicating in particular the type of repository (institutional, standard repository for the discipline, etc.). In case the dataset cannot be shared, the reasons for this should be mentioned (e.g. ethical, rules of personal data, intellectual property, commercial, privacy-related, security-related).

- ***Archiving and preservation (including storage and backup)***
  Description of the procedures that will be put in place for long-term preservation of the data. Indication of how long the data should be preserved, what is its approximated end volume, what the associated costs are and how these are planned to be covered.

6 Allocation of resources

The MONOCLE infrastructure has been designed as an open infrastructure from the start, therefore the effort and cost of making the data FAIR is part of the overall MONOCLE budget. It is the responsibility of each sensor provider within the project to ensure that their sensors adhere to the agreed standards, with support provided through Work Package 5 which is dedicated to data interoperability and accessibility. The development and maintenance of the MONOCLE backend are the responsibility of PML, who will continue to maintain access for at least three years beyond the end of the project.

7 Data security

To safeguard original data, backups will be made at the site where they are hosted. The nature of the MONOCLE data back-end is such that copies can be stored there, but this is not a requirement – it is designed to function both as a centralized and distributed data system.

Copies of data will, in general, not be backed-up at the MONOCLE back-end provided that they can be retrieved again from the source. The same applies to the use of Earth Observation and auxiliary data. Loss of such data would potentially cause delays due to the need to download them again from the source, but this will not be much different than restoring the data from tape backups.

A number of data repositories have been set up to safeguard specific project outputs, such as software, publications, sensitive data and frozen versions of sensor data. These will be accompanied by DOIs and are described in more detail in the document accompanying D9.3 “Open Data Repositories”.

8 Ethical aspects

Ethical aspects are mainly relevant for data gathered through citizen science initiatives. These data will be treated according to the ethics procedures laid out in D10.1, in summary these procedures cover the following aspects:

- Details on the procedures and criteria used to identify/recruit research participants
- Details on the informed consent procedures for the participation of humans.
- Templates of the informed consent forms
- Information sheets provided to participants
- Procedures regarding the recording of imagery where humans are identifiable

Appendix

Data Register Template

The example shows the information collected through the data register. Included are descriptions of the fields and an example covering Earth observation data.

Organisation	**Dataset reference & Name**		**Dataset description/outline**		**Standards & metadata**
Name of organisation providing the data. Also reference any other ownership, i.e. if you have bought commercial data and you have rights to use but must attribute etc	*A reference label. Should be unique when combined with your organisation name.*	*Simple description of the dataset, try to include as much information as possible on*	*Spatial Resolution & extent*	*Temporal resolution and extent*	*any standardised metadata that accompanies the dataset*
PML based on satellite data from ESA and NASA	CCI_reference_chlor_a	ESA OC-CCI archive consisting of global 4 x 4km ocean colour data. Consisting of individual RRS bands and derived chlor_a, the dataset has per pixel bias and rmsd uncertainty	resolution: 4km extent: -180,-90,180,90	1997-09-04T00:00:00.000Z to 2017-10-01T00:00:00.000Z	Files contain CF compliant metadata but currently no xml/ISO 19115 metadata exist

How will data be shared	software/protocol required for sharing	data access policy (open/locked/partial - give details, e.g. embargo time)	stored in MONOCLE Backend (yes/no - if no please say why/where it will be stored
list data services or custom websites	*list the protocols available for data access*	*data policy, such as groups that can use, whether it is only accessible to project partners or whether there is a time based embargo*	*whether the data will be stored in the MONOCLE back end or not. If not, describe how data will be stored.*
WMS: **https://vortices.npm.ac.uk/thredds/wms/CCI_ALL-v3.1-MONTHLY?service=WMS&version=1.3.0&request=GetCapabilities** **WCS**: **https://vortices.npm.ac.uk/thredds/wcs/CCI_ALL-v3.1-MONTHLY?service=WCS&version=1.1.0&request=GetCapabilities**	OGC WCS OGC WMS	Fully open data	No Archive will be proxied through the backend using the WMS/WCS links provided