

Project	SEAMLESS No 101004032	Deliverable	D1.3
Dissemination	Public	Nature	ORDP
Date	30 June 2021	Version	1.0



Deliverable 1.3

Initial data management plan

Deliverable Contributors:	Name	Organisation	Role / Title
Deliverable Leader	Stefano Ciavatta	PML	Project coordinator
Contributing Author(s)	Pierre Brasseur	UGA/CNRS	PI
	Laurent Bertino	NERSC	PI
	Gianpiero Cossarini	OGS	PI
	Lars Nerger	AWI	PI
Reviewer(s)	Jessica Heard	PML	Project manager
Final review and approval			



Project	SEAMLESS No 101004032	Deliverable	D1.3
Dissemination	Public	Nature	ORDP
Date	30 June 2021	Version	1.0

Document History:

Release	Date	Reason for Change	Status	Distribution
1.0	30/06/21	reviewed	Released	Public
0.0	22/06/21	Initial document	Released	Internal

To cite this document

This document is in a public deliverable. It should be cited as “Ciavatta S, Brasseur P, Bertino L, Cossarini G and L Nerger, 2021. Deliverable D1.3: Initial data management plan, EU H2020 SEAMLESS project, grant agreement No 101004032.



Project	SEAMLESS No 101004032	Deliverable	D1.3
Dissemination	Public	Nature	ORDP
Date	30 June 2021	Version	1.0

TABLE OF CONTENTS

1. Executive Summary	4
2. Scope	4
3. Data summary.....	4
4. FAIR data	6
4.1. Making data findable, including provisions for metadata	6
4.2. Making data openly accessible and interoperable	7
5. Allocation of resources.....	8
6. Data security	8
7. Ethical aspects	8
8. Appendix: Data Register Template	9



Project	SEAMLESS No 101004032	Deliverable	D1.3
Dissemination	Public	Nature	ORDP
Date	30 June 2021	Version	1.0

1. Executive Summary

SEAMLESS participates in the Open Research Data Pilot (ORDP) and follows the FAIR (Findable, Accessible, Interoperable, Re-usable) data paradigm and this is reflected in this Initial Data Management Plan (DMP).

This DMP details the management of data generated within the SEAMLESS project. It covers:

- what data will be generated
- how data will be handled during and after the project
- who owns a data set and with whom it is shared
- how sharing of data within and outside the project is organised
- what formats, metadata and standards the data will adhere to.

The SEAMLESS data sets consist of numerical simulation outputs generated by a range of marine ecosystem models, thus adhering to principles of accessibility and interoperability is important to ensure the data are useful throughout the project and to others. This DMP describes relevant procedures put in place at the start of the project.

Scope

This document, comprising the Data Management Plan, is intended for internal and external use, describing the mechanisms that SEAMLESS will put in place to ensure all public data follow the FAIR (Findable, Accessible, Interoperable, Re-usable) data management principles.

The Data Management Plan will be updated as a “live” document during the lifetime of the project, with two scheduled release dates. The current document presents the status and planning at month 6 of the three-year project.

2. Data summary

The overall objective of SEAMLESS is to provide the Copernicus Marine Service (CMEMS) with new capabilities to deliver indicators of climate-change impacts and food security in marine ecosystems.

To achieve this objective, SEAMLESS will produce data of marine physical and biogeochemical variables and indicators from model simulations. These simulations cover the European regional seas matching the domains of the Monitoring and Forecasting Centres (MFCs) of the Copernicus Marine Service, as well as the the North Atlantic Ocean (see Figure 1). In addition, one-dimensional simulations (water column profiles) will be delivered in correspondence of data-rich monitoring sites for comparison with field observations (Figure 1).



Project	SEAMLESS No 101004032	Deliverable	D1.3
Dissemination	Public	Nature	ORDP
Date	30 June 2021	Version	1.0

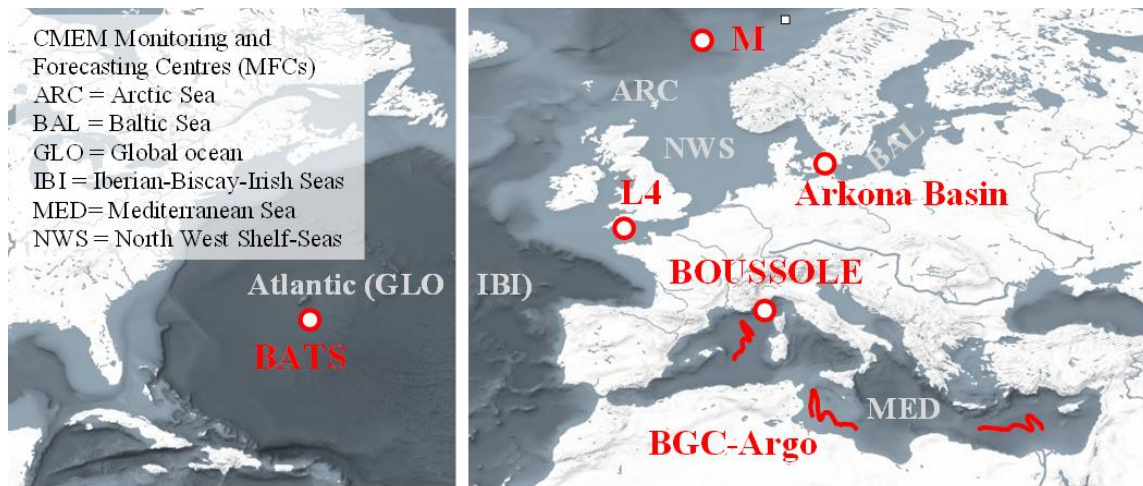


Figure 1. SEAMLESS will produce model output data for the six three-dimensional domains of the CMEMS Monitoring and Forecasting Centres (MFCs) listed and represented in the map. In addition, the project will deliver one-dimensional simulations (water column profile) in correspondence of data-rich monitoring sites highlighted by the red circles. The red lines represent the trajectories of the biogeochemical-Argo floats (BGC-Argo) simulated with Lagrangian configurations of the biogeochemical model of the Mediterranean Sea.

The project will produce model outputs with the primary aim of assessing the performance of the data assimilation methods and biogeochemical models developed or advanced in the project, through comparison with observations available in the literature or provided by project stakeholders.

In addition, we expect that the improved simulations of the biogeochemical variables and indicators can be exploited by project stakeholders to:

- 1) understand better the functioning of marine ecosystems (e.g. carbon export, phytoplankton community succession, productivity);
- 2) assess the health of the ecosystem in relation to, e.g., acidification, deoxygenation and eutrophication indicators;
- 3) evaluate alternative monitoring strategies (e.g. deployment of biogeochemical-Argo floats and/or gliders);
- 4) consider upgrades of the current modelling and assimilation systems of the CMEMS MFCs (e.g. using hybrid assimilation methods rather than variational only in the Mediterranean and North-West European Seas)

The outputs will include primarily all the ecosystem indicators that are targeted by SEAMLESS in all the simulated domain (see Table 1). Additional biogeochemical variables will be produced by the partners in relation to the observations available for model assessment in their target region (e.g. nutrients). Physical data will also be produced as by-products of the coupled biogeochemical-physical model simulations.



Project	SEAMLESS No 101004032	Deliverable	D1.3
Dissemination	Public	Nature	ORDP
Date	30 June 2021	Version	1.0

Table 1 List of biogeochemical **indicators** which will be delivered as primary output of the SEAMLESS model simulations

- 1) Particulate organic carbon (POC)
- 2) Trophic efficiency (as zooplankton normalized by phytoplankton concentrations)
- 3) Primary production
- 4) pH
- 5) Oxygen concentration
- 6) Phytoplankton functional types (PFT; carbon biomass)
- 7) Phytoplankton phenology (initiation, peak and duration of the blooms, as quantified by the chlorophyll concentrations)

The biogeochemical datasets produced by SEAMLESS are listed and described in Tables 2 relative to the CMEMS MFCs regions, and in Table 3 relative to the data-rich sites.

Table 2 List of the model output datasets that will be produced by SEAMLESS for the CMEMS MFC regions shown in Figure 1. The primary content of the datasets will be the biogeochemical indicators listed in Table 1.

The datasets will be produced as three-dimensional fields of daily or monthly values. All the datasets are in NetCDF format

Partner	Region	Model	Size of model output (Tb)
PML	NWS	NEMO-ERSEM	1.0
AWI	BAL	NEMO-ERGOM	0.5
IGE	GLO/ATL	NEMO-PISCES	2.0
NERSC	ARC	HYCOM-ECOSMO	0.05
OGS	MED	NEMO-BFM	0.5Tb per year simulated

Table 3 List of the model output datasets that will be produced by SEAMLESS for the data rich sites shown in Figure 1. The primary content of the datasets will be the biogeochemical indicators listed in Table 1.

The datasets will be produced as one-dimensional profiles of daily or monthly values. All the datasets are in NetCDF format

Partner	Region	Model	Size (Gb)
PML	NWS	GOTM-ERSEM	0.1
AWI	BAL	GOTM -ERGOM	0.1
IGE	GLO/ATL	GOTM -PISCES	0.1
NERSC	ARC	GOTM -ECOSMO	0.1
OGS	MED	GOTM -BFM	10

3. FAIR data

4. 1. Making data findable, including provisions for metadata

All SEAMLESS research data will be curated according to the 'FAIR' principle, i.e. to be Findable, Accessible, Interoperable and Re-usable. In the following, a short overview is given of the building blocks and guiding principles to reach this goal. As the SEAMLESS modelling and assimilation



This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 101004032.

Project	SEAMLESS No 101004032	Deliverable	D1.3
Dissemination	Public	Nature	ORDP
Date	30 June 2021	Version	1.0

systems are under active development, further detail will be added to future releases of this document as design decisions are made.

Any documentation such as model description and simulation set-up for final datasets will be linked within the data register (see Section 4.2) and a copy will be kept in the SEAMLESS project management portal. Documents for public dissemination will be made available on the project website.

By default, all data generated in SEAMLESS will be openly available (see Data Access, below), with the exception of unprocessed, uncorroborated data produced in test simulations 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 SEAMLESS is possible. Reference to existing FAIR data sources is to be preferred over duplication.

Initially, the metadata used to describe datasets made available will follow OGC standards. As a common ontology the CF conventions (cfconventions.org) will be followed or extended. These metadata conventions ensure that data are identifiable using appropriate search terms and key words.

4.2. Making data openly accessible and interoperable

Data generated as part of SEAMLESS will be free of cost to the user. 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 to which this specific user has been granted access.

The model output datasets produced within the project will be stored in the servers of the partners producing the data. The data will be made freely available to users, on request.

Model output data will be kept available for a minimum of three years after the end of SEAMLESS. 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.

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

An internal-use data register will be maintained, insofar as the data are not self-describing and catalogues in the services already mentioned. A snapshot will be created for each DMP release. A template is included in the Appendix. Datasets that meet the criteria for dissemination have not yet been generated for this version of the DMP.



Project	SEAMLESS No 101004032	Deliverable	D1.3
Dissemination	Public	Nature	ORDP
Date	30 June 2021	Version	1.0

4. Allocation of resources

The development and maintenance of the SEAMLESS project management portal is the responsibility of PML, who will continue to maintain access for at least three years beyond the end of the project.

Management of the repositories of the datasets in the partners' servers is responsibility of the partners.

5. Data security

To safeguard the final model output datasets, backups will be made at the site where they are hosted.

Copies of test simulation data will, in general, not be backed-up by SEAMLESS after the final datasets are delivered.

A number of data repositories have been set up to safeguard specific project outputs, such as software (such as the PML GitHub repository), publications (such as the Plymouth Marine Science Electronic Archive, PlymSEA), sensitive data (internal project management databases, which are backed up onto a separate PML server plus an external additional backup copy).

6. Ethical aspects

Ethical aspects are mainly relevant for data of a personal nature. These data will be treated according to the ethics procedures laid out in the Ethics section of the GA, in summary these procedures cover the following aspects:

- Details on the informed consent procedures for the participation of humans
- Templates of the informed consent forms
- Templates of the user and stakeholder questionnaires.



Project	SEAMLESS No 101004032	Deliverable	D1.3
Dissemination	Public	Nature	ORDP
Date	30 June 2021	Version	1.0

7. Appendix: Data Register Template

The example shows the information collected through the data register. Included are descriptions of the fields and an example covering the model output datasets for the North West European Shelf-Seas domain.

Organisation	Dataset reference & Name	Dataset description/outline			Standards & metadata	How will data be shared	Software/protocol required for sharing	Data access policy (open/locked/partial - give details, e.g. embargo time)	Location of storage	Documentation (deliverable)
<i>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</i>	<i>A unique reference label</i>	<i>Simple description of the dataset, try to include as much information as possible</i>	<i>Spatial Resolution & extent</i>	<i>Temporal resolution and extent</i>	<i>Any standardised metadata that accompanies the dataset</i>	<i>List data services or custom websites</i>	<i>List the protocols available for data access</i>	<i>Data policy, such as groups that can use, whether it is only accessible to project partners or whether there is a time based embargo</i>	<i>Where and how data will be stored</i>	
Example: PML	REA_NWS_2019 (Reanalysis 2019 North-West Shelf seas domain)	NEMO-ERSEM output dataset for physical and biogeochemical variables	CMEMS MFC N domain, 4-dimensional 7km horizontal resolution, variable vertical resolution	Daily means	NetCDF-CF	Transferred to user on request	PML ftp server	open	Data and back-up are stored in the PML servers	D6.1



