Project Title	Blue-Cloud 2026: A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters
Project Acronym	Blue-Cloud 2026
Project Number	101094227
Type of project	RIA – Research and Innovation Action
Topics	HORIZON-INFRA-2022-EOSC-01
Starting date of Project	01 January 2023
Duration of the project	42 months
Website	www.blue-cloud.org

D2.2 – New Blue Data Infrastructures – Service Analysis Report

Work Package	WP2 | FAIR compliant Discovery and Access services for marine domains & beyond
Task	T2.2 | Horizontal expansion: federation of additional Blue Data Infrastructures
Lead author	Dick M.A. Schaap (MARIS)
Contributors	Enrico Boldrini (CNR-IIA), Roberto Roncella (CNR-IIA), Peter Thijsse (MARIS), Paul Weerheim (MARIS), Serge vd Horst (MARIS), Robin Kooyman (MARIS), Bert Broeren (MARIS), Tjerk Krijger (MARIS), Guy Cochrane (EML-EBI), Stéphane Pesant (EMBL-EBI), Rob Finn (EMBL-EBI), Suran Jayathilaka (EMBL-EBI), Lili Meszaros (EMBL-EBI), Antonio Novellino (ETT), Marco Alba (ETT), Øystein Godøy (SIOS), Heikki Lihavainen (SIOS), Ilkka Matero (SIOS), Ivan Rodero (EMSO), Raul Bardaji (EMSO)
Peer reviewers	Laura Beranzoli (INGV), Marina Tonani (MOI)
Version	V1.0
Due Date	31/10/2023
Submission Date	29/11/2023

Dissemination Level

X	PU: Public
	CO: Confidential, only for members of the consortium (including the Commission)
	EU-RES. Classified Information: RESTREINT UE (Commission Decision 2005/444/EC)
	EU-CON. Classified Information: CONFIDENTIEL UE (Commission Decision 2005/444/EC)
	EU-SEC. Classified Information: SECRET UE (Commission Decision 2005/444/EC)

Funded by the European Union
Blue-Cloud2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227

Version History

Revision	Date	Editors	Comments
0.1	30/10/2023	Dick M.A. Schaap (MARIS) and contributors	First draft
0.2	23/11/2023	Dick M.A. Schaap (MARIS)	Final draft after processing internal reviews from Marina Tonani (MOI) and Laura Beranzoli (INGV)
0.3	29/11/2023	Sara Pittonet (Trust-IT), Pasquale Pagano (CNR)	Quality check
1.0	29/11/2023	Pasquale Pagano (CNR)	Final version for submission

Keywords

Blue-Cloud 2026, EOSC, Marine research, Open Science; Data Discovery; Data Federation; Data Access; Federated Discovery; Federated Access

Disclaimer

The Blue-Cloud 2026 project has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101094227. Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

Glossary of terms

Item	Description
API	Application Programming Interface
BDI	Blue Data Infrastructure
CDI	Common Data Index (SeaDataNet)
CMEMS	Copernicus Marine Environment Monitoring Service
CMC INSTAC	Copernicus Marine In Situ Assembly Centre
CSW	Catalogue Service for the Web (OGC standard)
DAB	Discovery and Access Broker
DD&AS	Data Discovery & Access Service
ELIXIR-ENA	European Nucleotide Archive (ELIXIR service)
EMODnet	European Marine Observation and Data network
EMSO	European Multidisciplinary Seafloor and water column Observatory
EOSC	European Open Science Cloud
ERDDAP	Environmental Research Division's Data Access Program
EuroGOOS	European contribution to GOOS program
GDAC	Global Data Assembly Centre
GLOSS	Global Sea Level Observing Service
GNSS	Global Navigation Satellite System
GOOS	Global Ocean Observing System
GUI	Graphical User Interface
INSDC	International Nucleotide Sequence Database Collaboration
ISO	International Organization for Standardization
JSON	JavaScript Object Notation
MGnify	ELIXIR-MGnify platform facilitates the assembly, analysis and archiving of microbiome-derived nucleic acid sequences
NODC	National Oceanographic Data Centre
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OGC	Open Geospatial Consortium
PSMSL	Permanent Service for Mean Sea Level
pyCSW	OGC API and OGC CSW server, written in Python
SeaDataNet	pan-European network for marine and ocean data management
SIOS	Svalbard Integrated *Arctic* Earth Observing System
SQL	Structured Query Language
TSC	Technical and Scientific Committee
UI	User Interface
VLab	Virtual Laboratory
VRE	Virtual Research Environment

Funded by the European Union
Blue-Cloud2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227

Item	Description
WEkEO	WEkEO is one of the 5 Copernicus DIAS, bringing in the CMEMS, C3S and CAMS
WFS	Web Feature Service (OGC standard)
WMS	Web Map Service (OGC standard)
WMTS	Web Map Tile Service (OGC standard)
XML	Extensible Markup Language

EXECUTIVE SUMMARY

The **Blue Cloud Data Discovery and Access Service** (**DD&AS**) is one of the core components of the Blue-Cloud technical framework. It It facilitates discovery and retrieval of marine data sets and data products from blue data infrastructures (BDIs), such as key EU infrastructures such as Copernicus Marine Service, EMODnet, SeaDataNet, and more, for external users in stand-alone mode. It interacts with the Blue-Cloud Virtual Research Environment, the component federating computing platforms and analytical services, for populating the VRE data pool. As part of the predecessor pilot Blue-Cloud project, the first operational release of the Blue-Cloud DD&AS has been deployed.
The current version of the service is federating in total 8 BDIs and with 9 data services. As part of the Blue-Cloud 2026 project, it is planned to expand the DD&AS federation with data discovery and access services from 4 new BDIs, namely:

- European Multidisciplinary Seafloor and water column Observatory (EMSO)
- Svalbard Integrated *Arctic* Earth Observing System (SIOS)
- ELIXIR - MGnify
- EMODnet Physics.

The implementation of the Blue Cloud DD&AS largely depends on machine-to-machine interactions between central components of the DD&AS and web services / APIs as managed and operated by the BDIs. In this deliverable a short overview is given of the current architecture and functionality of the Blue-Cloud DD&AS to provide good insight in the DD&AS conceptual principles. Per each new BDI a description is given of their organisation, structure, aims, contents, data discovery and access services, and available web services and/or APIs, followed by an assessment of how each BDI might be federated in the DD&AS, fitting into the overall concept. The gathering of information about each new BDI and the follow-up analyses and assessment have been performed thanks to a two-way dialogue between MARIS and CNR-IIA, as core DD&AS developers, with each of the new BDIs and associated technical providers. This activity started at the plenary WP2 working session at the Project Kick-Off meeting, on 13–15 February 2023 in Pisa – Italy, and continued during the the first Technical and Scientific Committee (TSC) meeting, 28-29 March 2023 in Amsterdam – Netherlands. Since then, the managers of the new BDIs have compiled information about their BDI and services, following a provided template. Bilateral web conferences followed with managers of each of the BDIs during August – October 2023 to discuss findings and suggestions for the federation.

This consultation phase resulted in concrete plans for federating each of the 4 new BDIs, which are documented in this Deliverable **D2.2 - New Blue Data Infrastructures – Service Analysis Report**. These plans will be worked out and taken into development in the coming period as part of the WP2 activities.

Funded by the European Union
Blue-Cloud2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

Funded by the European Union
Blue-Cloud2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227

1. Introduction

The term Blue Data Infrastructures (BDIs) refers to those infrastructures that observe and/or aggregate marine and ocean observation data from other sources, and that make these data available for users via discovery and access services as part of their service offer.

The pilot Blue-Cloud project (2019-2022) deployed the first operational release of the **Blue Cloud Data Discovery and Access Service** (**DD&AS**), accessible at: https://data.blue-cloud.org/search. The current DD&AS service is federating 8 Blue Data Infrastructures (BDIs) with 9 data discovery and access services. It provides a common user interface for discovery and retrieval of multi-disciplinary data sets as managed by each of the federated BDIs. Users are able to select and download selected data sets.

As part of the Blue-Cloud 2026 project, activities are planned for further optimising the services of the DD&AS. As a starting point, an analysis of the existing 8 BDIs was undertaken by MARIS and CNR-IIA together with the managers of the given infrastructures, which have resulted in a list of planned activities reported in **Deliverable D2.1** - **Existing DD&AS and Blue Data Infrastructures – Review and Specifications for Optimisation Report.**

The Blue-Cloud2026 project is now expanding the DD&AS federation with 4 new BDIs, namely EMSO, SIOS, ELIXIR-MGnify, and EMODnet Physics. This task was started at the Blue-Cloud2026 project kick-off meeting, on 13 – 15 February 2023, Pisa – Italy, where an overview was given of the current DD&AS and its underlying concept for federation of BDIs. A more in-depth follow-up took place at the Blue-Cloud2026 Technical and Scientific Committee (TSC) meeting, 28 – 29 March 2023, Amsterdam – The Netherlands, where targets and options for optimisation of the DD&AS were discussed, and where also opportunity was given to the new BDIs to present themselves. As a follow-up to the TSC meeting, the technical representatives of the new BDIS were requested by the WP2 leader (MARIS) to describe their infrastructures, and in particular their data discovery and access mechanisms, in a document based on a provided template. This request was met by all representatives and their draft descriptions were analysed by MARIS and CNR-IIA, with a focus on how the services of the new BDIs might (be made) fit in the federation concept of the DD&AS, adopting the same principles as already applied for the 8 BDIs already included in the DD&AS. This analysis was followed by bilateral meetings involving MARIS, CNR-IIA and representatives of each BDI to validate technical and content aspects, and to discuss the way forward towards federation. Most of these meetings were organised as online sessions, while a physical meeting took place at EBI in Hinxton – UK in order to discuss with the large group of both the existing ELIXIR-ENA and the new ELIXIR-MGnify BDIs.

The insights gained about the new BDIs and how to federate their services in the Blue-Cloud DD&AS have provided the input for this Deliverable **D2.2 - New Blue Data Infrastructures – Service Analysis Report**.

Funded by the European Union
Blue-Cloud2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227

This is a working document and it includes still a number of open questions which need to be worked out in collaboration with technical and content representatives of the BDIs and the core developers of the Blue-Cloud DD&AS, i.e. MARIS and CNR-IIA. Therefore, this document also includes a number of actions which should lead to further precising and elaborating the technical and contents solutions for integrating the data discovery & access services of the 4 new BDIs into the DD&AS.

Funded by the European Union
Blue-Cloud2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227

2. Overall set-up of DD&AS

In the following paragraph a short summary is given of the overall concept and set-up of the DD&AS including those principles and aspects which are relevant for federating additional BDIs.

2.1. Overall concept

The federated Data Discovery & Access Service (DD&AS) provides users with an easy and FAIR service for discovery and access to multi-disciplinary data sets and data products managed and provided by leading Blue Data Infrastructures (BDIs). The federation facilitates sharing of datasets as input for analytical and visualisation services and applications, that are hosted and further developed as VLabs and WorkBenches in the Blue-Cloud VRE. The DD&AS has been developed, is operated, and is being upgraded and expanded by MARIS together with CNR-IIA and CINECA (member of EUDAT), interacting with each of the BDIs. The currently federated BDIs and characterisation of their data resources are given in the following figure.

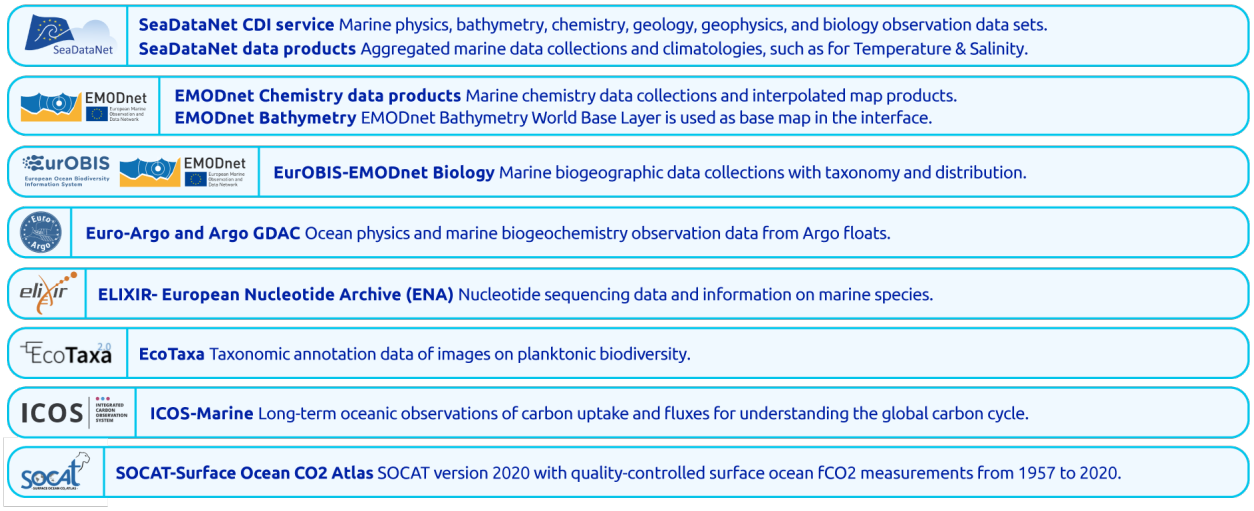


Figure 1 Current Blue Data Infrastructures (BDIs) federated in the DD&AS

The DD&AS facilitates common discovery and access to more than 10 million marine datasets for physics, chemistry, geology, bathymetry, biology, biodiversity, and genomics. The DD&AS works with **brokerage services both at metadata and data level**.

Discovery and selection are done in a two-step approach - from **data collections** using a common metadata profile, to detailed **granular records** using extended metadata profiles- and fully based on web services (such as **OGC-CSW, OAI-PMH, ERDDAP, DCAT, and dedicated APIs**), as published and maintained by connected BDIs:

- The first step has a focus on identifying interesting data at an aggregated collection level, with free search, geographic and temporal criteria as main query operators;
- The second step has a focus on drilling down within identified collections and their BDIs to get more specific data at granular level, using again free search, geographic and temporal criteria, but this time at a granular level, concerning individual observation data sets, and including additional search criteria which are specific per BDI;
- Finally, users are facilitated by a shopping mechanism to download and store the retrieved data collections on their own machines or in a data pool as part of the Blue-Cloud VRE.

For the **metadata brokerage at the first level of data collections**, CNR-IIA has advanced and deployed an internal service, namely a Blue-Cloud discovery broker service based on their DAB technology. This middleware harvests metadata at collection level from each of the BDIs, using their existing web services or APIs. The DAB service transforms the harvested XML files from each of the BDIs into a common ISO Blue-Cloud collection profile, which is published by the DAB service by means of a Blue-Cloud OGC-CSW service with a common XML profile for each BDI. See image below.

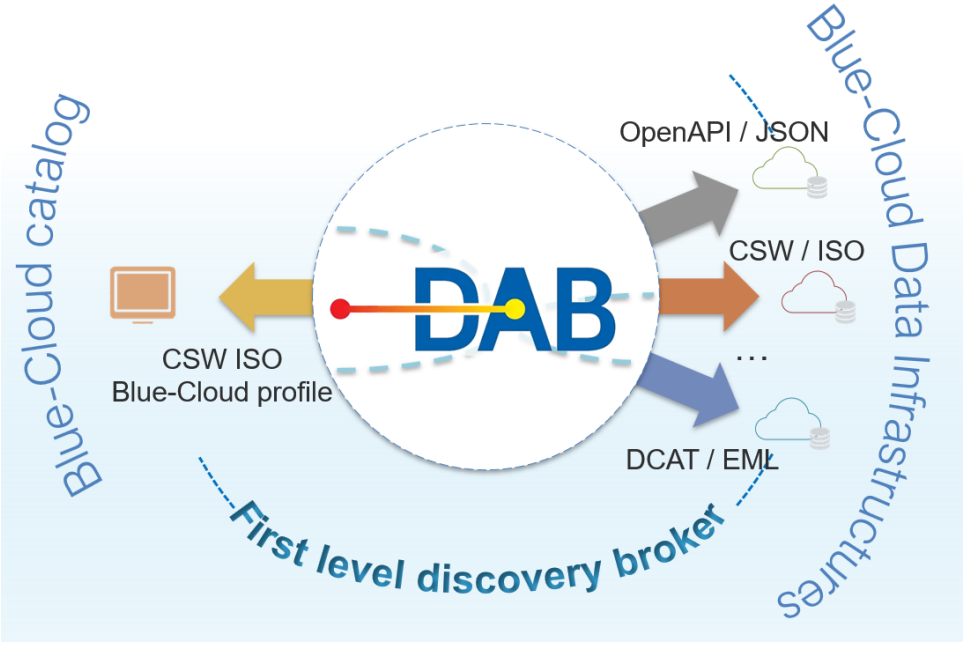


Figure 2 Blue-Cloud first level discovery broker component harmonizes the protocols and data models published by different heterogeneous BDIs to a harmonized CSW service based on ISO 19115

The returned records are expressed according to the Blue-Cloud collection metadata profile, an ISO 19115 based metadata profile encoded using the recent ISO 19115-3:2016 XML schema implementation. In total

13 metadata elements from ISO 19115 are considered as the common elements of the profile, as they are deemed to be the more useful for discovery of Blue-Cloud collections. The common Blue-Cloud metadata elements are:

- IDENTIFIER: Blue-Cloud unique and persistent code for the metadata record;
- TITLE: a characteristic, and often unique, name by which the collection is known;
- ABSTRACT: a short description of the collection;
- KEYWORD: a commonly used word, formalised word or phrase used to describe the subject;
- BOUNDING_BOX: extent of the resource in the geographic space given as a bounding box;
- TEMPORAL_EXTENT: time period covered by the content of the collection;
- PARAMETER: name of the attribute described by the measurement value;
- INSTRUMENT: measuring instrument used to acquire the data;
- PLATFORM: platform from which the data were taken;
- ORGANIZATION: organization associated with the collection;
- DATESTAMP: the latest update date of the metadata description;
- REVISION_DATE: the latest update date of the data;
- RESOURCE_LINKS: download links where available and useful.

The DAB service of CNR-IIA regularly harvests and thus updates the output of the Blue-Cloud CSW ISO v. 2.0.2 service per BDI. Currently, this is done on a weekly basis. CSW is a well-known standard web service of the Open Geospatial Consortium (OGC), recommended by many initiatives for sharing metadata on the web**.**

MARIS then harvests these common formatted XML entries on a regular basis from each of the Blue-Cloud **OGC CSW services** and integrates these into a SQL database which is indexed with Elastic Search. This processing makes full free text searching very efficient and fast. This way, the first step of the Blue-Cloud query process is powered, which has been integrated by MARIS in the interface of the Blue-Cloud Data DD&AS. And the common metadata base also serves the second level of the Blue-Cloud query process and the retrieval of data download links, but then only for the BDIs with one level of data. All has been set up as an automatic process without human intervention, driving weekly updating from the connected BDIs and synchronisation from the DAB CSW services to the indexed Blue-Cloud catalogue service as part of the Blue-Cloud DD&AS.

The **second step** drills down within identified collections to get more specific data, using free search, geographic and temporal criteria, but this time at granular level, and including additional BDI-specific search criteria. The **Blue-Cloud Data brokerage service,** operated by MARIS, performs the master role in the Blue-Cloud DD&AS. Regularly, it retrieves the latest Blue-Cloud level 1 metadata catalogue from the Blue-Cloud metadata brokerage, and ingests this into the discovery interface, whereby users can query

the catalogue at level 1. The common level 1 metadata catalogue includes sufficient metadata for each BDI to allow the first level queries at collection level with a few selection criteria and this way to identify which of the BDIs holds interesting data sets. The Blue-Cloud level 1 metadata catalogue should also contain sufficient additional metadata to allow more specific searching at level 2 for those BDIs that only have data collections and other data products, but no service at granular level. While for other BDIs, supporting deeper searching at level 2 – granular level -, customised search profiles have been formulated, which allow the data broker to interact with the provided web services and APIs of the BDIs.

The Blue-Cloud Data brokerage service also contains a shopping mechanism with basket and ledger, by which users (external users and VRE) and BDIs can be informed about shopping transactions and their status in time. It interacts with the Blue-Cloud Data Cache to give it precise instructions about retrieving data sets from the BDIs and to insert these for temporary storage, and to bundle these as downloadable data packages for each shopping order. It interacts with the Marine-ID service as users need to login to submit shopping baskets and to have access to the transaction ledger. It interacts with registered users and VRE to inform and instruct them about data packages that are ready for downloading by users or retrieval for ingestion by the VRE. Finally, it also interacts with the Blue-Cloud Data Cache to receive information about the actual downloading by users and retrieval for ingestion by the VRE in order to update the ledger.

More information about the architecture of the current Blue-Cloud DD&AS can be found in the Deliverable D2.8 – Blue-Cloud Architecture (release 3) from the preceding Blue-Cloud pilot project and which is available at: https://zenodo.org/records/10065882

3. Description and analysis per BDI

Preliminary technical analyses and assessments have been made in dialogue between MARIS and CNR-IIA as core DD&AS developers together with contactpersons of each of the 4 new BDIs and possible associated technical providers. This started at the plenary WP2 working sessions at the Project Kick-Off meeting, 13–15 February 2023 in Pisa – Italy, and the first Technical and Scientific Committee (TSC) meeting, 28-29 March 2023 in Amsterdam – Netherlands. At these meetings the common principles for optimisation of the DD&AS were discussed, which can be summarised as:

- The services of the BDI should facilitate machine-to-machine operations
- The services of the BDI should facilitate discovery and access to data sets by downloading
- The services should have a TRL level > 7 and be managed by the BDI as operational services, which will be sustained for longer term
- The services should be documented, detailing functionality, data model, and queries
- The services should make use of controlled vocabularies
- The services should be able to give access to open data, even if part of the data might be restricted

Following the TSC meeting, managers of the 4 new BDIs were requested to detail their BDI and services, adopting a given template. These descriptions have been delivered and after analyses a number of bilateral meetings have taken place between the core developers, MARIS and CNR-IIA, and each of the 4 new BDIs. In the discussions, the principles, as listed above, have been taken onboard. These bilateral meetings took place in the period from August to October 2023 and were dedicated to review the provided BDI details and to discuss how the web services of the BDIs could become part of the DD&AS federation, fitting the two level approach, going from collections to granular records. These meetings have resulted in refinement of the BDI descriptions as earlier given in the provided templates and these form the basis for the following BDI descriptions.

Funded by the European Union
Blue-Cloud2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227

3.1. European Multidisciplinary Seafloor and Water Column Observatory (EMSO ERIC)

The European Multidisciplinary Seafloor and Water Column Observatory, European Research Infrastructure Consortium (EMSO ERIC) is a distributed research infrastructure consisting of ocean observation systems across the European seas. It serves as a distributed research infrastructure and stores a wide range of data for marine monitoring purposes. The main objectives of EMSO ERIC are to observe, perceive, and understand phenomena occurring on the seafloor and the water column over temporal scales that range from seconds to decades. To achieve this, EMSO ERIC deploys scientific instruments with physicochemical and biological sensors and develops specific tools to understand the obtained data. Initiated in 2008, the EMSO ERIC consortium has grown thanks to the European Union's support and its members' commitment. Now, the consortium comprises 27 institutions from Europe and has 11 observatory sites with permanent instrumentation and 3 test sites where the research centers can deploy instrumentation prototypes.

EMSO ERIC develops and promotes common standards, vocabularies, software tools, and services for ocean data management, which are freely available from its portal and are widely adopted and utilized. Furthermore, EMSO ERIC maintains and publishes all data obtained from the observatory sites.

3.1.1. Data discovery and access service component

As a distributed research infrastructure, EMSO ERIC's consortium members distribute and manage metadata and data according to their established workflows and EMSO ERIC specifications. Each member can maintain data traceability and exert control over it. Despite this decentralized system, it is also possible to access data from all institutions and observatories through a single web service, a **federated ERRDAP**, facilitating broad and unified access. The EMSO ERIC central management office operates the federated ERDDAP service.

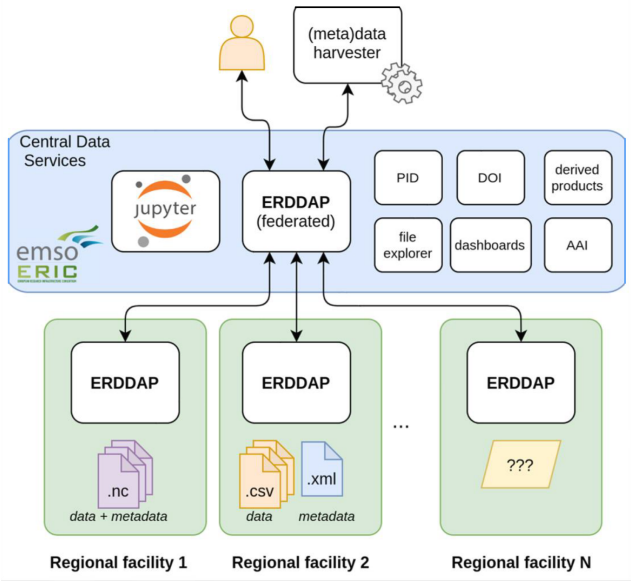


Figure 3 Architecture of the EMSO data management and distribution as a federated ERRDAP system (source EMSO)

This approach is being deployed to all data observatory sites and also expanding the range of observation parameters. The federated ERDDAP service is already operational, steadily increasing its data offerings, while the number of connected data observatory sites and range of included variables is expanding. This way, EMSO ERIC is making and has already made significant progress in optimizing the functioning of its data discovery and access service by establishing a unified metadata specification approach and delivery which is broadly accepted and an exchange mechanism based on machine-to-machine interfaces.

3.1.1.1 Name and web address

The EMSO ERIC federated ERDDAP service can be found at:

https://erddap.emso.eu

3.1.1.2 Types and number of data sets and/or data products

Currently, the datasets available on ERDDAP are mainly time series obtained from fixed stations. Some of these stations have sensors at different depths. The data from these time series can be collected in real-time or in “delayed mode”. Therefore, many of these datasets are constantly being updated. Through ERDDAP filters, users can create subsets of the datasets for a specific time and parameter.

Figure 4 Geographic overview of the EMSO Observation sites

The EMSO ERIC federated ERDDAP service currently exposes at least 157 entries, delivering datasets generated by the instruments installed at EMSO observation sites as depicted in the image above.

3.1.1.3 Discovery and access mechanisms - how does it function

The data discovery and access are based on ERDDAP. Additionally, the EMSO ERIC data portal provides pointers to ERDDAP servers and aims at providing other functionalities and value-added services on top of ERDDAP. Next to discovery and access this could include other functionality as provided by ERDDAP such as subsetting and graphics, while also added-value functionality might be developed by EMSO. ERDDAP is versatile and robust, able to handle various types of data. Numerous institutions have adopted ERDDAP worldwide, in particular for exchange of operational oceanography observations, and as recommended by GOOS.

In the context of EMSO ERIC, ERDDAP is a platform for accessing a wide range of datasets spanning multiple disciplines and regions, thereby fostering research collaboration and scientific discovery. The

data discovery process in EMSO ERIC is designed to be as intuitive and streamlined as possible. Users can search for data using a variety of parameters, such as the dataset's title, the data's time range, geographic location, or the parameters the data represents. Moreover, ERDDAP provides a clear, standardized metadata description for each dataset. This aids researchers in finding relevant datasets quickly and efficiently, irrespective of their size or complexity. Moreover, ERDDAP offers a user-friendly graphical interface, which allows users to visually explore the available datasets, assisting in locating the appropriate data for their research needs.

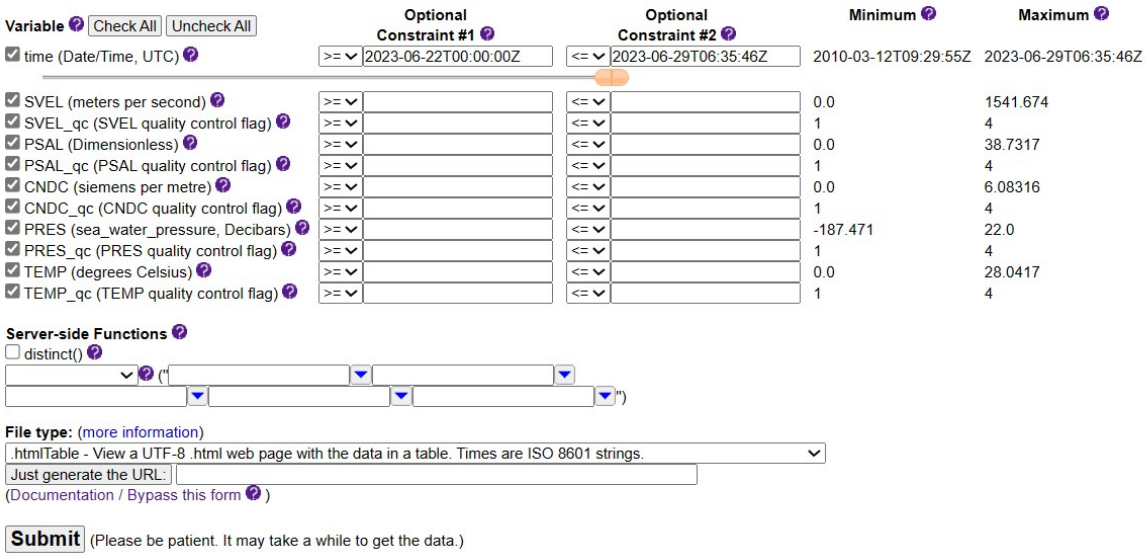


Figure 5 ERDDAP user interface

Once the data has been discovered, ERDDAP offers versatile **data access** mechanisms, enabling users to download data in various standard data formats such as .csv, .nc (NetCDF), .json, .mat (MATLAB), .xls (Excel), and many others. Furthermore, ERDDAP supports OPeNDAP, allowing users to remotely access parts of a dataset without the need to download the entire file.

Users can request data subsets, specifying the exact parameters, time range, and geographic area they are interested in. ERDDAP then processes these requests and generates the desired output file, which can be downloaded directly or accessed via a URL for future reference. This way, the data management system saves on bandwidth and storage, providing researchers with only the data they need. This also implicates, that ERRDAP supports both retrieving pre-defined data sets by their metadata records as well as making subsets or slices from those data sets. Furthermore, the **ERDDAP's RESTful API** makes

programmatically accessing and downloading data straightforward, allowing researchers to integrate EMSO ERIC data access directly into their data analysis workflows.

The data discovery and access mechanisms in EMSO ERIC data management also emphasize **interoperability**. ERDDAP supports CF (Climate and Forecast) conventions and COARDS (Cooperative Ocean/Atmosphere Research Data Service) conventions, making the data easily interpretable and compatible with a wide array of software. This promotes the use of the data across multiple platforms and applications, facilitating collaboration and data sharing among researchers worldwide (i.e., approximately 100 organizations in at least 17 countries based on the ERDDAP public documentation).

3.1.1.4 Any new developments underway

The integration of datasets into ERDDAP is an ongoing concern, which strives at establishing local ERDDAP configurations for inclusion of data streams from all EMSO Observation sites and expanding the range of data types covered. Moreover, the EMSO data portal is under continuous improvements, mainly in backend functionalities and value-added products and services. For the federation in the Blue-Cloud 2026 Data Discovery & Access Service (DD&AS) it is guaranteed, that the EMSO federated ERDDAP service will be operated as a robust service with increasing structured metadata and data provision, so that it will function as a reliable and persistent node for the Blue-Cloud federation.

3.1.2 Standards in use

3.1.2.1 Metadata format(s) - short overview and references to detailed documentation

The EMSO ERIC metadata format is based on the “OceanSITES Data Format reference Manual v1.4”. However, specific adaptations have been incorporated to adapt them to the EMSO ERIC infrastructure and needs. At the heart of the EMSO ERIC metadata structure is a set of general conventions. These conventions guide the inclusion and presentation of all attributes within a dataset. While all attribute names are expected to be present, optional attributes may bear an empty string as their value. For instances where multiple values are essential, a semicolon-separated list is utilized. An example of such an instance is as follows:

- "project": "EMSODEV; EMSO-Link; ENVRI-FAIR"

In contrast, for optional attributes, a structure may look like this:

- "required_attribute": "very important metadata",
- "optional_attribute": ""

Every EMSO-compliant dataset in ERDDAP is expected to incorporate a set of **Global Attributes**. The following table shows the Global Attributes:

Table 1 – EMSO Global attributes

Global Attributes	Description	Required	Multiple
Conventions	Conventions used in the dataset (e.g. OceanSITES, ACDD, etc.)	Yes	Yes
date_created	Creation date	Yes	Yes
institution_edmo_code	EDMO code of the creator's organization	Yes	Yes
institution_edmo_uri	URI pointint to the EDMO page de of the creator's organization	Yes	Yes
geospatial_lat_min	The southernmost latitude, a value between -90 and 90 degrees	Yes	No
geospatial_lat_max	The northernmost latitude, a value between -90 and 90 degrees	Yes	No
geospatial_lon_min	The westernmost longitude between 180 and 180	Yes	No
geospatial_lon_max	The easternmost longitude between 180 and 180	Yes	No
geospatial_vertical_min	Minimum depth of measurements in metres (negative for above sea level)	Yes	No
geospatial_vertical_max	Maximum depth of measurements in metres (negative for above sea level)	Yes	No
time_coverage_start	Start date of the data in UTC	Yes	No
time_coverage_end	End date of the data in UTC	Yes	No
update_interval	Update interval for the file (following ISO 8601). If not applicable Use “void”	Yes	No
site_code	EMSO site code of the platform	Yes	Yes
emso_facility	EMSO facility name	No	Yes
Source	Platform type name, should be a L06 prefered label (prefLabel)	No	No
platform_code	OceanSITES platform code (leave blank if it does not exist)	No	Yes

wmo_platform_code	WMO platfrom code (leave blank if it does not exist)	No	Yes
data_type	Type of data, in most cases 'OceanSITES data time-series data'	No	No
format_version	OceanSITES format version	No	No
network	List of the networks	Yes	Yes
data_mode	Data mode from OceanSITES table 4, possible values are R, P D or M	No	No
title	Free-format text describing the dataset, for use by human readers	Yes	No
summary	Longer free-format text describing the dataset	Yes	No
keywords	Please use 'SeaDataNet Parameter Discovery Vocabulary'	No	Yes
keywords_vocabulary	URI of the keywords vocabulary used	No	No
project	Acronyms of the projects funding the dataset	No	Yes
principal_investigator	Name of the principal investigator	Yes	Yes
principal_investigator_email	Email of the principal investigator	Yes	Yes
doi	Digital Object Identifier (DOI) of the dataset	No	No
license	License name (SPDX short identifier), use of CC-BY-4.0 is strongly recommended.	Yes	No
license_uri	URI pointing to an SPDX license, use of CC-BY-4.0 is strongly recommended	Yes	No

Data variables also contain some **variable attributes** that need to be included as metadata. The following table shows the variable attributes:

Table 2 EMSO variable attributes

Global Attributes	**Description**	**Required**	**Multiple**
standard_name	Climate and Forecast (CF) standard name (P07 vocabulary)	Yes	No
units	Variable units, should be the preferred label from a P06 definition	Yes	No

comment	Free-text to add comments on the variable	No	No
coordinates	Variable coordinates/dimensions, usually "TIME; DEPTH; LATITUDE; LONGITUDE; SENSOR_ID"	Yes	Yes
ancillary_variables	Related variables, e.g. quality control flags, standard deviations, etc.	No	Yes
_FillValue	Fill value	No	Yes
reference_scale	Reference scale of the variable (e.g. ITS90 for temperature)	No	No
sdn_parameter_name	Variable name (should be the preferred label from the P01 term)	Yes	No
sdn_parameter_urn	Variable code (should be an identifier from P01)	Yes	No
sdn_parameter_uri	URI for the P01 term	No	No
sdn_uom_name	Variable units, should be the preferred label from a P06 definition	Yes	No
sdn_uom_urn	Units identifier from SeaDataNet P06 vocabulary	Yes	No
sdn_uom_uri	Units URI from SeaDataNet P06 vocabulary	No	No
sensor_model	Sensor model (L22 preferred label)	Yes	Yes
sensor_SeaVoX_L22_code	Sensor model (L22 identifier)	Yes	Yes
sensor_reference	Sensor model (L22 URI)	Yes	Yes
sensor_manufacturer	Sensor model, L35 preferred label	Yes	Yes
sensor_manufacturer_uri	Sensor model (URI from the L35 term)	Yes	Yes
sensor_manufacturer_urn	Sensor model (should be preferred label from the L35 term)	Yes	Yes
sensor_serial_number	Unique identifier for the sensor	Yes	Yes
sensor_mount	One of the possible sensor mounts (see the sensor mount list)	Yes	Yes
sensor_orientation	One of the possible sensor orientations (see the sensor orientation list)	Yes	Yes

Although the “ancillary_variables” term is not required, it is mandatory to set it in case there is a quality control column related to the parameter.

The possible sensor mounts are auto-explained and are the following:

- mounted_on_fixed_structure
- mounted_on_surface_buoy
- mounted_on_mooring_line
- mounted_on_bottom_lander
- mounted_on_moored_profiler
- mounted_on_glider
- mounted_on_shipborne_fixed
- mounted_on_shipborne_profiler
- mounted_on_seafloor_structure
- mounted_on_benthic_node
- mounted_on_benthic_crawler
- mounted_on_surface_buoy_tether
- mounted_on_seafloor_structure_riser
- mounted_on_fixed_subsurface_vertical_profile

The possible sensor orientations are the following:

- downward: For example, an ADCP measuring currents from its location to bottom.
- Upward: For example, an ADCP measuring currents towards the surface.
- Horizontal: For example, Optical sensor looking 'sideways' from the mooring line or on a ship CTD frame.

Additionally, the **dimension** (time, latitude, longitude, depth) **attributes** are also required to be compliant with EMSO metadata specification:

Table 3 EMSO dimension attributes

Dimension Attributes	Description	Required	Multiple
long_name	Human-readable label for the variable	Yes	No
standard_name	Climate and Forecast (CF) standard name (P07 vocabulary)	Yes	No
units	Variable units, should be the preferred label from a P06 definition	Yes	No
comment	Free-text to add comments on the variable	No	No
ancillary_variables	Related variables, e.g. quality control flags, standard deviations, etc.	No	Yes
_FillValue	Fill value	No	Yes

Funded by the European Union
Blue-Cloud2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227

sdn_parameter_name	Variable name (should be the preferred label from the P01 term)	Yes	No
sdn_parameter_urn	Variable code (should be an identifier from P01)	Yes	No
sdn_parameter_uri	URI for the P01 term	No	No
sdn_uom_name	Variable units, should be the preferred label from a P06 definition	Yes	No
sdn_uom_urn	Units identifier from SeaDataNet P06 vocabulary	Yes	No
sdn_uom_uri	Units URI from SeaDataNet P06 vocabulary	No	No

Finally, the **Quality Control (QC) attributes** are also required:

Table 4 EMSO Quality Control attributes

Dimension Attributes	Description	Required	Multiple
long_name	Human-readable label, it is suggested to use the parameter long_label and add "quality control flags" at the end	Yes	No
conventions	OceanSITES QC Flags	Yes	No
flag_values	QC value	Yes	Yes
flag_meanings	Meaning of the QC Flag	Yes	Yes

The QC flag value and meaning must be one of the following:

- 0: unknown
- 1: good_data
- 2: probably_good_data
- 3: potentially_correctable_bad_data
- 4: bad_data
- 5: nominal_value
- 8: interpolated_value
- 9: missing_value

3.1.2.2 Data format(s) - short overview and references to detailed documentation

EMSO ERIC provides data from the observatories through ERDDAP and supports various formats. The following data formats are available in ERDDAP:

- NetCDF

- CSV and ASCII
- JSON
- Other formats: .mat, .xls, .png, .pdf

Each data format has its peculiarities and preferred use. For example, the NetCDF format is especially useful for handling multidimensional array data, commonly used in scientific and geographical applications. On the other hand, the ASCII format is widely used for text data easily readable by humans and machines. ERDDAP provides a robust set of tools and functionalities for the conversion between these formats. In addition, ERDDAP's built-in functionality allows users to extract subsets of data in their preferred format, simplifying the manipulation and analysis of the data. In addition to these common formats, ERDDAP supports several special formats for specific data types, such as high-frequency radar data in NetCDF4 format. ERDDAP provides transparent access to these data formats through its web interface, allowing users to choose the format that best suits their needs and applications. For more details on ERDDAP's data formats, the detailed ERDDAP documentation is publicly available at https://coastwatch.pfeg.noaa.gov/erddap/download/setup.html, which provides a comprehensive view of how to interact with these data formats.

3.1.2.3 Use of controlled vocabularies - which, where, how

EMSO ERIC data makes use of several controlled vocabularies from the NERC Vocabulary Service (https://vocab.nerc.ac.uk/):

Table 5 Controlled Vocabularies as used by EMSO

Which	Where (in metadata information)	URL
EDMO	SeaDataNet -European Database of Marine Organizations Codes	https://edmo.seadatanet.org/
ROR	Research Organization Registry	https://ror.org/
P01	Parameter vocabulary	http://vocab.nerc.ac.uk/collection/P01/current/
P02	Parameter codes	http://vocab.nerc.ac.uk/collection/P02/current/
P06	Units	http://vocab.nerc.ac.uk/collection/P06/current/
L05	Sensor types	http://vocab.nerc.ac.uk/collection/L05/current/
L06	Platform types	http://vocab.nerc.ac.uk/collection/L06
L22	Sensor model	http://vocab.nerc.ac.uk/collection/L22/current/

Funded by the European Union
Blue-Cloud2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227

3.1.2.4 Data access policy - if yes, which and how deployed, using any AAI

EMSO ERIC provides open access to its data without a specific EMSO access policy. However, the data is distributed under the Creative Commons Attribution license (CC-BY), which permits and encourages its unrestricted use and distribution. Access logs are used for user tracking and enhancing the user experience.

The use of authentication and authorization infrastructure is limited to EMSO ERIC internal services. The EMSO ERIC privacy policy is available at:

https://emso.eu/privacy-policy/

3.1.3 Your suggestions for aggregating data sets as collections in order to fit the two-step approach of the Blue-Cloud DD&AS, going from collections to granules

The current output of the EMSO federated ERDDAP service is organised in collections, which already should fit the Blue-Cloud level 1 as it is. A further grouping into subcollections might be feasible, for example:

- **Thematic:** Collections can be organized according to corresponding themes such as physical oceanography data (temperature, salinity, current speed, etc.), biological oceanography data (chlorophyll, phytoplankton, etc.), or chemical oceanography data (pH, nutrients, dissolved oxygen, etc.).
- **Geographic:** Data could be geographically arranged, grouping data by specific ocean regions, such as the North Atlantic, Mediterranean Sea, etc.
- **Temporal**: Another method could be time-based data grouping like annual or monthly collections or specific seasonal data.
- **Expedition/Campaign**: Another way to group data could be based on specific expeditions or campaigns.

However, most straight forward could be to make use of the collections as already made available through the ERDDAP service. Currently, these comprise 150+ records, which will increase steadily, but still at reasonable numbers.

The collection records could be broken down into granular data sets. This functionality is provided by the ERDDAP interface, which allows to query and retrieve a subset like an individual observation, a time series, a depth in a profile, etc. The level of granularity will depend on the nature of the data and the analysis requirements. For instance, for physical oceanography data, a granule could represent a single measurement of temperature or salinity at a specific location and time. For biological oceanography data, a granule could be a specific chlorophyll measurement or the count of a specific type of phytoplankton in

a defined water sample. ERDDAP provides the ability to subset and rescale the data, allowing access to granular data sets efficiently. This can be particularly useful when dealing with large datasets that would be challenging to handle in their entirety.

3.1.4 Any data subsetting services in use or under development – URLs, function, how to operate

The harmonization process is ongoing, and several EMSO ERIC facilities have made their datasets available in the federated ERDDAP services. ERDDAP has functionality for subsetting, although its performance decreases in case of many large files.

3.1.5 Hosting environment

The EMSO ERIC central management office provides the federated ERDDAP service in a cloudbased environment. EMSO ERIC facilities operating ERDDAP servers are listed below:

- Universitat Politecnica de Catalunya
- National Institute of Oceanography and Applied Geophysics - OGS, Division of Oceanography
- CNRS
- Ifremer
- Laboratoire Oceanographique de Villefranche
- National Research Council of Italy
- Marine Institute

Other facilities will make their ERDDAP servers available shortly.

3.1.6 Organisational aspects

3.1.6.1 Main operator(s); data providers

The EMSO ERIC central management office operates the federated ERDDAP server. The main operators and data providers are the same for the EMSO ERIC facilities:

- Universitat Politecnica de Catalunya
- National Institute of Oceanography and Applied Geophysics - OGS, Division of Oceanography
- CNRS
- Ifremer
- Laboratoire Oceanographique de Villefranche
- National Research Council of Italy

- Marine Institute

3.1.6.2 Contact details for technical developments

raul.bardaji@emso-eu.org and ivan.rodero@emso-eu.org

3.1.7 Conclusions for Data Infrastructure

The focus for the Blue-Cloud DD&AS federation will be on the output of the EMSO federated ERDDAP service at:

https://erddap.emso.eu

The ERDDAP output is organised by EMSO in collections and currently there are 150+ records. This is regularly increasing as EMSO is further deploying the ERDDAP federation scheme to all EMSO observation sites, while also expanding the range of data types and range of instruments. EMSO is undertaking The federated ERDDAP service is well maintained by EMSO as it is considered as their operational service for management and giving access to the data sets as acquired through EMSO.

The metadata and data formats as used by EMSO are well documented and being adopted for all sites. When comparing with the Blue-Cloud DAB broker metadata profile (see $2.1), it appears that the 'EMSO Global Attributes' already cover content for most of the required Blue-Cloud metadata fields tags (title, abstract, keywords, bounding box, temporal extent, platform, organisation) while the 'EMSO Variable Attributes' provide additional detailed content for the other Blue-Cloud core metadata fields (parameters, instruments). This will facilitate the mapping of the EMSO output to the level 1 Blue-Cloud common metadata profile. Moreover, the EMSO ERDDAP output contains several additional attributes which might be used to expand the Blue-Cloud metadata set.

Considering level 2, further analysis is needed how to make use of the ERDDAP subsetting ability for querying and retrieving data sets, associated to the collections at level 1. ERDDAP provides the functionality to query and retrieve a subset like an individual observation, a time series, a depth in a profile, etc. The level of granularity could depend on the nature of the data and the analysis requirements. This can be particularly useful when dealing with large datasets that would be challenging to handle in their entirety. A first approach could be to review the feasibility to keep Blue-Cloud level 2 also at collection level, only adding extra criteria for refining the collection search. If downloadable files become too large to handle, then a further step could be to add subsetting criteria, that should be overall applicable to all collection records, in order to break the large data sets down in subsets.

Another aspect is semantic interoperability. EMSO makes use of several controlled vocabularies from the SeaDataNet - NERC Vocabulary Service and EDMO (see $3.1.2.3), thereby not only including the literal

description, but also the URIs and URNs. This implies that the EMSO ERDDAP output already is anticipating the semantic interoperability requirements that are part of the planned optimisation of the Blue-Cloud DD&AS as documented in Blue-Cloud 2026 deliverable D2.1. This aspect will be further analysed during the mapping exercise for the Blue-Cloud level 1 and level 2. EMSO already seems to conform to the proposed Blue-Cloud principle to include and provide 'triples' for each term, consisting of literal term, code of term (URN), and URL of associated vocabulary (URI). Having these triples onboard of the EMSO output and the Blue-Cloud DAB broker will facilitate deploying the semantic brokerage as planned in deliverable D2.1.

All data distributed by EMSO ERIC is **Open**, but the specific policy depends on the regional infrastructure. By default, CC-BY is used. There is work in progress at EMSO for deciding on a unified policy for the data distributed by EMSO ERIC via ERDDAP.

Funded by the European Union
Blue-Cloud2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227

3.2. Svalbard Integrated Arctic Earth Observing System (SIOS)

3.2.1. Data discovery and access service component

The **Svalbard Integrated Arctic Earth Observing System (SIOS)** is a collaborative interdisciplinary effort to develop and maintain a regional observational system for long-term measurements in and around Svalbard, addressing Earth System Science questions related to Global Change. The observing system and research facilities offered by SIOS build on the extensive observation capacity and diverse world-class research infrastructure provided by many institutions already established in Svalbard. This includes a substantial capability for utilising remote sensing resources to complement ground-based observations. From this solid foundation, SIOS envisions a significant contribution to the systematic development of new methods and observational design in Svalbard. This knowledge can advance other observational networks in the Arctic and elsewhere.

SIOS is aiming at more efficient use and better integration of the observing system based on a distributed data management system, an open access program that includes logistical support, as well as training and education activities. Working groups, task forces and other SIOS components pursue these aims in direct and structured dialogue with scientists, user groups, policy-makers and other porters of societal and scientific needs. SIOS brings observations together into a sustained, coherent and integrated observational programme. SIOS focuses on processes and their interactions between the different spheres, i.e., biosphere, geosphere, atmosphere, cryosphere, and hydrosphere. The core observational programme of SIOS provides the research community with systematic observations that are sustained over time, yet dynamic enough to be adapted as new methods appear or society poses new questions.

3.2.1.1 Name and web address

Svalbard Integrated Arctic Earth Observing System:

https://sios-svalbard.org/

3.2.1.2 Types and number of data sets and/or data products

SIOS is an interdisciplinary effort and covers many types of data and disciplines, ranging from in situ measurements to remote sensing and model output. The core observational component is the **SIOS Core Data** which are in situ measurements.

Currently the SIOS Data catalogue holds **502389 datasets**, not all of these are in a harmonised form and some come from third party sources but are relevant for SIOS activities and region of interest. The number of **SIOS Core Data** is not directly available in the catalogue as not all datasets are properly tagged yet and

some are not accepted yet due to pending harmonisation efforts. However, there are currently more than **54 core datasets** that are properly tagged and their number is increasing

3.2.1.3 Discovery and access mechanisms - how does it function

There is a human interface to the data catalogue available at:

https://sios-svalbard.org/metsis/search?f%5B0%5D=dataset_level%3ALevel-1

In the human interface, the user can distinguish between parent and child records (a parent record can have many child records). This is the interface to the central catalogue which is merely a discovery metadata catalogue containing information that is harvested from partner data repositories (and third party catalogues) using nightly incremental harvests and a full harvest every 3 months. For harmonised data the human interface has visualisation and transformation services when data are served as CF-NetCDF over OPeNDAP or through OGC WMS.

In addition there is a machine interface to the unified catalogue. This is based on the pyCSW software and supports multiple discovery metadata standards (e.g. GCMD DIF, ISO19115, INSPIRE) and protocols (e.g. OAI-PMH, OGC CSW). The OAI-PMH service is better for incremental harvesting; however, at regular intervals a full harvesting might be required as the OAI-PMH service does not support deletions of records. See $3.2.3 for the endpoints of the web services.

3.2.1.4 Any new developments underway

There is a strong focus on the harmonisation of SIOS Core Data and services on top of this. The vocabulary server is undergoing upgrading to facilitate mappings between e.g. CF standard names and GCMD Science Keywords. In addition the machine interface to the catalogue is being upgraded and the visualisation software for in situ data is being upgraded.

3.2.2 Standards in use

3.2.2.1 Metadata format(s) - short overview and references to detailed documentation

The internal format for the discovery metadata is MMD[1]. Information from partner data repositories is harvested, converted to MMD and ingested in the SIOS data catalogue. Harvesting of information from partner or third-party sources is supported using various flavours of GCMD DIF and ISO19115 (using GCMD Science Keywords and OSGeo protocol descriptions). Support for harvesting information from schema.org and DCAT is in progress. EML is supported but yet not in active use for automated harvesting, only manual (i.e. selected datasets where there is no other possibility). In the human interface to the data catalogue

[1] *https://htmlpreview.github.io/?https://github.com/metno/mmd/blob/master/doc/mmd-specification.html*

discovery metadata for data records in the catalogue can be downloaded as ISO19115, GCMD DIF (multiple flavours of both) as well as INSPIRE and WMO profiles of ISO19115 and the native MMD format. In the machine interface, discovery metadata content is provided in ISO19115 (Inspire and WMO profiles) or GCMD DIF in addition to Dublin Core (using pyCSW[2] as the software provisioner). The ISO19115 output is less homogeneous as it has no prescribed vocabularies.

Concerning use metadata (explaining how to interpret datasets) Climate and Forecast Conventions and Darwin Core is actively used.

3.2.2.2 Data format(s) - short overview and references to detailed documentation

For SIOS Core Data the following is required:

- CF-NetCDF 1.6 or higher (discrete sampling geometries, one station per file, allowing aggregation in time).
- Darwin Core Archives.

3.2.2.3 Use of controlled vocabularies - which, where, how

GCMD Science Keywords, CF Standard Names, OSGeo protocol descriptions and MMD native vocabularies are used. MMD native vocabularies are published on the web as SKOS Concepts[3]. The human interface of the data catalogue is providing direct links to this vocabulary server for all the controlled vocabularies used in MMD. For CF standard names, links to the NERC vocabulary server[4] are in place on the interface. Support for GCMD Locations and GCMD providers, as well as GEMET Inspire keywords is included in the MMD model.

The search model is based on GCMD Science Keywords and mappings are used wherever possible. Protocol descriptions of GCMD or OSGeo is used to determine the purpose of links found in the discovery metadata and enabling of services or other functionality.

3.2.2.4 Data access policy - if yes, which and how deployed, using any AAI

The SIOS data policy can be found at:

https://sios-svalbard.org/sites/sios-svalbard.org/files/common/SIOS_Data_Policy.pdf

No AAI is enforced, but user registration/AAI is required to access higher order services like the basket or transformation services. The system allows users to log in using several mechanisms. The built in solution

[2] *https://pycsw.org/*
[3] *https://www.w3.org/TR/skos-reference/*
[4] *https://vocab.nerc.ac.uk/standard_name/*

and federated log in using e.g. ORCID, Edugain, Google, Microsoft. Federated AAI is not enforced currently and there are no plans to require this for straightforward data discovery in the human interface, only for higher order services.

3.2.3 Web services and API's - URLs, function, how to operate

Except for the search API (based on pyCSW, offering OAI-PMH, OGC CSW and OpenSearch) all other data discovery API's (i.e. SolR) are internal to the system for the time being. The whole architecture is service oriented where Drupal is used as an interactive frontend to various API's for visualisation, transformation etc. The architecture relies on OPeNDAP (for data access and visualisation of non gridded datasets through self developed APIs), OAI-PMH/OGC CSW (for data discovery) and OGC WMS (for visualisation of gridded datasets).

OGC WPS is currently being used for transformation services for gridded data which are built into the interactive search interface, but this is in the process of being replaced by OGC API Processes. Transformation services for gridded data are thus disabled for the time being. For non gridded data data transformation is supported through a self defined API based on FastAPI. These services are currently not available to external users, only for the interactive interface.

All open APIs are fully open to anyone that would like to use them. Information on whether a dataset is served through OPeNDAP or OGC WMS is provided in the discovery metadata that can be extracted using the pycsw end point. The relevant end points for SIOS are:

- OGC CSW: https://sios.csw.met.no
- OAI-PMH: https://sios.csw.met.no/csw?mode=oaipmh&verb=Identify
- OpenSearch: https://sios.csw.met.no/csw?mode=opensearch&service=CSW&version=3.0.0&request=GetCapabilities

OAI-PMH and OGC CSW are well tested.

3.2.4 Your suggestions for aggregating data sets as collections in order to fit the two-step approach of the Blue-Cloud DD&AS, going from collections to granules

Use either OAI-PMH or OGC CSW to discover datasets and their web services and rely on OpeNDAP for combining data.

3.2.5 Any data subsetting services in use or under development – URLs, function, how to operate

Transformation of data, including subsetting in variable, temporal or spatial domains as well as reformatting of data is done relying on OpeNDAP and dedicated web services. These web services are, as indicated above, under reimplementation and not available to external users for the time being. This information can be provided at a later time.

3.2.6 Hosting environment

Central services are run in the operational infrastructure of the Norwegian Meteorological Institute. This relies on existing data centres operated by the partners. The central node runs on a number of virtual machines orchestrated using OpenStack, with some backend processing services running in a HPC system. Everything in a high availability framework. K8S is available, but has not been utilised yet as the existing environment has proven more stable over time.

3.2.7 Organisational aspects

3.2.7.1 Main operator(s); data providers

SIOS currently has 15 involved organisations. Moreover, SIOS has many 3rd party providers, from which information and data sets are harvested and bundled in the SIOS output.

3.2.7.2 Contact details for technical developments

The following contact persons:

Øystein Godøy, o.godoy@sios-svalbard.org

Lara Ferrighi. l.ferrighi@met.no

Ilkka Matero, ilkka.matero@sios-svalbard.org

3.2.8 Conclusions for Data Infrastructure

The focus for the Blue-Cloud DD&AS federation will be on the output of the SIOS PyCSW based OAI-PMH service of SIOS at:

https://sios.csw.met.no/csw?mode=oaipmh&verb=Identify

While for the metadata mapping analysis, the focus will be on the GCMD DIF profile as it is supporting controlled vocabularies. Also, OAI-PMH seems to be better fit for incremental harvesting which might make sense in case of the high number of records.

A review is required considering which contents to take onboard. Blue-Cloud could focus only on the SIOS Core Data sets which are currently a low number (54 records) or go for the larger volume which currently

counts more than 500.000 records, originating fom 15 SIOS data providers and from 3rd parties. This larger contents will be further reviewed, taking into account a number of possible criteria, like:

- Only data from the SIOS data providers
- Only records that give access to data sets for downloading

By focusing on the records, provided by the SIOS providers, MARIS and CNR-IIA could see how each of their contributions might be reorganised in collections versus granular records, thereby taking into account the large amount of records, and the fact that the SIOS data providers in principle are participating to the Blue-Cloud 2026 project, which will allow for discussing and deploying possible changes, if required.

Once, the target set of metadata records has been concluded, in dialogue with SIOS, MARIS and CNR-IIA will then analyse deeper the metadata and data formats as used by SIOS, checking the available documentation and trying out services. This will include a mapping analysis of the SIOS GCMD DIF profile with the Blue-Cloud DAB broker common metadata profile (see §2.1), and options for expanding the Blue-Cloud metadata set with additional metadata fields.

Another aspect is semantic interoperability. SIOS makes use of a number of controlled vocabularies (see §3.2.2.3). Also, it is mentioned that SIOS has its own vocabulary service. It should be reviewed if vocabulary terms are not only included with their literal description, but also with the coding of the terms and the URL of the associated vocabularies. This will be required to fulfill the semantic interoperability requirements that are part of the planned optimisation of the Blue-Cloud DD&AS as documented in Blue-Cloud 2026 deliverable D2.1.

All data distributed by SIOS is **Open** and there is a SIOS data policy. No AAI is enforced, but user registration/AAI is required to access higher order services like the basket or transformation services.

3.3. EMODnet Physics

3.3.1. Data discovery and access service component

EMODnet Physics is one of the seven thematic lots of the European Marine Observation and Data network (EMODnet) that has successfully designed and deployed operational services providing ocean physical data and data products. EMODnet-Physics is an upstream ocean data integrating service and it is federating discovery and access from a number of source data infrastructures. It provides a single point of access to in situ ocean physics time-series data and vertical profiles, data products and metadata built with common standards, free of charge and no restrictions.

EMODnet Physics addresses the challenge of ensuring open access to harmonized integrated data across a very diverse range of ocean observing "systems", applying a "systems of systems" approach, and encompassing multi-scale and multi-platform/sensor observations. It has a a common procedure to access ocean physics parameters and products from several sources: available products are collections of in situ data, reanalysis and trends from in situ data, elaboration in space and/or time of in situ data and model output for a given parameter. Data products delivery mode ranges from real-time, near real time to validated long term time series.

EMODnet Physics does not run any observation platform itself, but metadata and data from key European oceanographic repositories and marine infrastructures (EuroGOOS, CMC INSTAC (Copernicus), SeaDataNet NODCs) are integrated with other available physics data sources such as the ICES database, PANGAEA repository, the Permanent Service for Mean Sea Level (PSMSL), the SONEL - GNSS data assembly centre for the Global Sea Level, the Global Sea Level Observing Service (GLOSS), European Multidisciplinary Seafloor and water column Observatory (EMSO), and others, to provide the most comprehensive in situ ocean physics data catalogue.

The available parameters cover temperature, salinity and currents profiles, sea level trends, wave height and period, wind speed and direction, water turbidity (light attenuation), underwater noise, river flow, and sea-ice coverage. EMODnet Physics products range from in situ data collections (time-series, profiles and datasets) as recorded by platforms (tide gauge, river stations, CTDs, etc.), to reanalysis, trends, and climatology. Data and data products are accompanied by metadata providing the user with information on what, where, when, who, etc. as well as quality check applied procedures. EMODnet Physics also implements standardized machine-to-machine procedures (ISO 19115-2/19139, OpenDAP, Web Coverage Service (WCS), Web Map Service (WMS), Web Feature Service (WFS), etc.).

Datasets from the various data sources (which can make available data in different formats – csv, txt, netcdf etc - and by means of different distribution tools – FTP, http download, webREST, etc) are linked in and, EMODnet Physics applies routines to harmonise data and metadata - it applies common vocabularies to complete metadata, applies a common data transport format, etc - and makes data ready for its stakeholders.

3.3.1.1 Name and web address

EMODnet Physics is operating its own back-office system with databases and a number of web services. These services are integrated with the EMODnet Central Portal (https://emodnet.ec.europa.eu/en) which features the user interfaces. The EMODnet Central portal provides a central map viewer:

https://emodnet.ec.europa.eu/en

and a central product catalogue:

https://emodnet.ec.europa.eu/geonetwork/srv/eng/catalog.search#/home

Which include maps and data product descriptions as provided by EMODnet Physics.

The driving services of EMODnet Physics are:

Geonetwork Catalogue:

Staging	Production
https://geonetwork.emodnet-physics.eu	https://prod-geonetwork.emodnet-physics.eu/geonetwork/

GeoServer:

Staging	Production
https://geoserver.emodnet-physics.eu	https://prod-geoserver.emodnet-physics.eu/geoserver/

ERDDAP:

Staging	Production
https://erddap.emodnet-physics.eu https://erddap.emodnet-physics.eu/ncMWS	https://prod-erddap.emodnet-physics.eu/erddap/ https://prod-erddap.emodnet-physics.eu/ncWMS/

For the federation in the Blue-Cloud DD&AS, the focus should be at the ERDDAP services.

3.3.1.2 Types and number of data sets and/or data products

The available parameters cover temperature, salinity and currents profiles, sea level trends, wave height and period, wind speed and direction, water turbidity (light attenuation), underwater noise, river flow, and sea-ice coverage. Data products are collections of in-situ data, reanalysis and trends of parameters, space and time aggregated in situ data and model outputs.

3.3.1.3 Platforms Types and data management

Operational oceanography deals with two types of data. First there are (near) real-time data required for operations at sea and daily and weekly forecasting activities. Second there are delayed-mode data when

the observed data are subject to further quality control; these data are particularly valuable for reanalysis work and to assist seasonal forecasting and climate monitoring and prediction where long-term stability is essential. Some platforms, such as gliders or research cruises, may not transmit (all) the data in real time to the Technical Assembly Centre. In such case, there might be an operational data stream that includes a subset of parameters delivered in near real time, while the complete dataset is recovered at the end of the mission (i.e. delayed mode). Other delayed mode observations derive from the reprocessing of near real time observations that undergo more exhaustive and advanced quality check procedures. The main types of in-situ observing systems as included in EMODnet Physics comprise the following.

- Argo - free-drifting profiling floats
- Buoys – drifting buoys, ice buoys, moored buoys and OCEANSITES
- Tide Gauges
- Gliders
- Marine Mammals CTD
- High Frequency Radars
- Ferry Box
- Research Vessels
- Voluntary Ship programs

3.3.1.4 Discovery and access mechanisms - how does it function

3.3.1.4.1 Map Viewer

The first tool to discover the EMODnet Physics data is the EMODnet Central Portal Mapviewer: https://emodnet.ec.europa.eu/geoviewer/

By selecting the “EMODnet Physics” as part of the “Catalogue”, a user can find relevant maps.

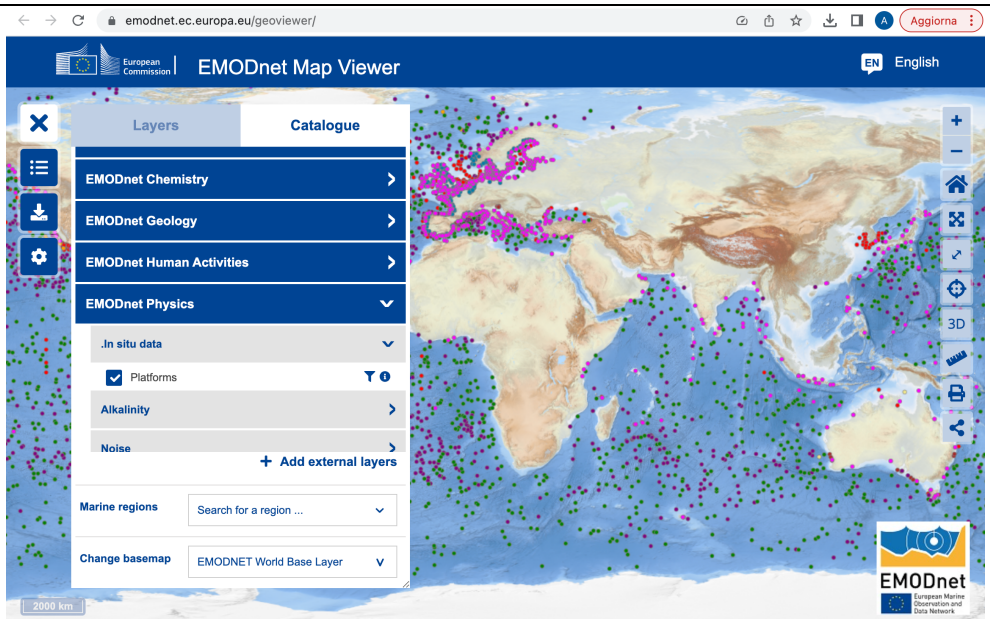


Figure 6 Map Viewer – Access to EMODnet Physics map layers

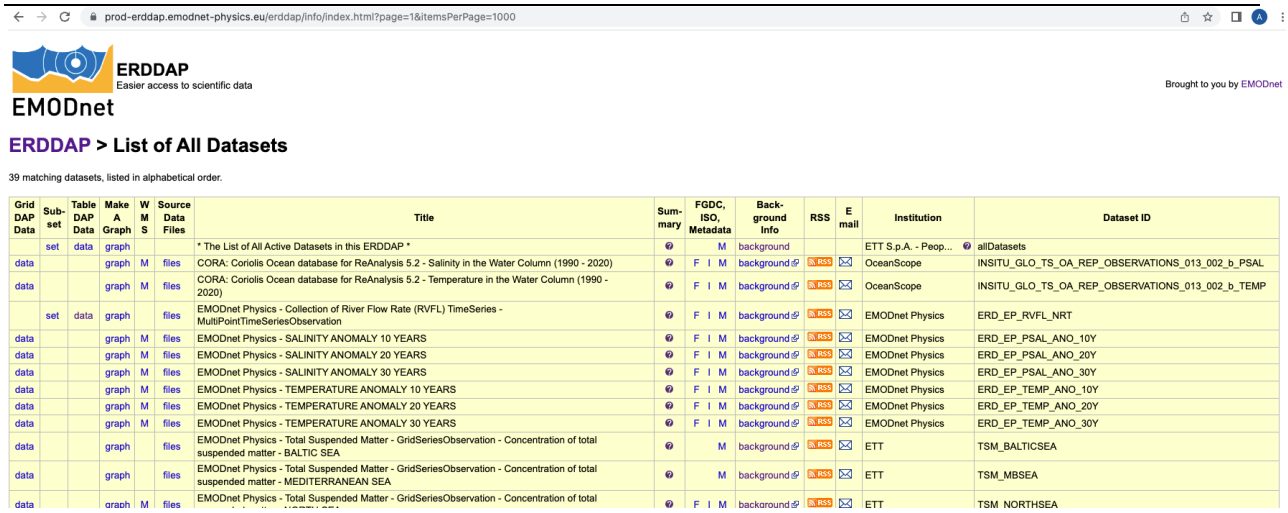
The system presents the platform that have made available at least one dataset during the last 7 days, in their last position. The users then can activate a "filter menu" to modify the query and have a fine tuning of the selection. In addition, the user can retrieve a platform page. Once loaded, the user can explore all the parameters as observed at the selected platform. Notably each platform has its unique url-id:

https://map.emodnet-physics.eu/platformpage/?platformid=3111018

The platform page is also the entry point to get links to the API of the selected platform.

3.3.1.4.2 ERDDAP API

ERDDAP is an Apache based data server that offers an easy and consistent way to download subsets of gridded and tabular scientific datasets in common file formats and make graphs and maps.



Grid DAP Data	Sub-set	Table DAP Data	Make A Graph	W M S	Source Data Files	Title	Sum-mary	FGDC, ISO, Metadata	Back-ground Info	RSS	E mail	Institution	Dataset ID
	set	data	graph			* The List of All Active Datasets in this ERDDAP *		M	background			ETT S.p.A. - Peop...	allDatasets
data			graph	M	files	CORA: Coriolis Ocean database for ReAnalysis 5.2 - Salinity in the Water Column (1990 - 2020)		F I M	background	RSS		OceanScope	INSITU_GLO_TS_OA_REP_OBSERVATIONS_013_002_b_PSAL
data			graph	M	files	CORA: Coriolis Ocean database for ReAnalysis 5.2 - Temperature in the Water Column (1990 - 2020)		F I M	background	RSS		OceanScope	INSITU_GLO_TS_OA_REP_OBSERVATIONS_013_002_b_TEMP
	set	data	graph		files	EMODnet Physics - Collection of River Flow Rate (RVFL) TimeSeries - MultiPointTimeSeriesObservation		F I M	background	RSS		EMODnet Physics	ERD_EP_RVFL_NRT
data			graph	M	files	EMODnet Physics - SALINITY ANOMALY 10 YEARS		F I M	background	RSS		EMODnet Physics	ERD_EP_PSAL_ANO_10Y
data			graph	M	files	EMODnet Physics - SALINITY ANOMALY 20 YEARS		F I M	background	RSS		EMODnet Physics	ERD_EP_PSAL_ANO_20Y
data			graph	M	files	EMODnet Physics - SALINITY ANOMALY 30 YEARS		F I M	background	RSS		EMODnet Physics	ERD_EP_PSAL_ANO_30Y
data			graph	M	files	EMODnet Physics - TEMPERATURE ANOMALY 10 YEARS		F I M	background	RSS		EMODnet Physics	ERD_EP_TEMP_ANO_10Y
data			graph	M	files	EMODnet Physics - TEMPERATURE ANOMALY 20 YEARS		F I M	background	RSS		EMODnet Physics	ERD_EP_TEMP_ANO_20Y
data			graph	M	files	EMODnet Physics - TEMPERATURE ANOMALY 30 YEARS		F I M	background	RSS		EMODnet Physics	ERD_EP_TEMP_ANO_30Y
data			graph		files	EMODnet Physics - Total Suspended Matter - GridSeriesObservation - Concentration of total suspended matter - BALTIC SEA		M	background	RSS		ETT	TSM_BALTICSEA
data			graph		files	EMODnet Physics - Total Suspended Matter - GridSeriesObservation - Concentration of total suspended matter - MEDITERRANEAN SEA		M	background	RSS		ETT	TSM_MBSEA
data			graph	M	files	EMODnet Physics - Total Suspended Matter - GridSeriesObservation - Concentration of total		F I M	background	RSS		ETT	TSM_NORTHSEA

Figure 7 ERDDAP interface of EMODnet Physics

To list the available datasets:

https://prod-erddap.emodnet-physics.eu/erddap/info/index.html?page=1&itemsPerPage=1000

ERDDAP supports both gridded data (under the ncWMS) and timeseries. Datasets with timeseries have "TableDAP" link to data (e.g. "EMODnet Physics - Collection of River Flow Rate (RVFL) TimeSeries – MultiPointTimeSeriesObservation"). User can query timeseries data from the ERDDAP Dataset ID page:

https://prod-erddap.emodnet-physics.eu/erddap/tabledap/ERD_EP_RVFL_NRT.html

The Dataset ID page is a web tool to construct and refine queries, e.g. **to list the operational platforms** (and their positions) in the dataset, the user has to select EP_PLATFORM_ID, LAT, LONG (and platform_code) and e.g. a Time

The query is:

https://erddap.emodnet-physics.eu/erddap/tabledap/EP_ERD_INT_RVFL_AL_TS_NRT.htmlTable?EP_PLATFORM_ID%2Clatitude%2Clongitude%2Cplatform_code&time%3E=2022-07-30T00%3A00%3A00Z&distinct()

to save the results as a csv

https://erddap.emodnet-physics.eu/erddap/tabledap/EP_ERD_INT_RVFL_AL_TS_NRT.csv?EP_PLATFORM_ID%2Clatitude%2Clongitude%2Cplatform_code&time%3E=2022-07-30T00%3A00%3A00Z&distinct()

EP_PLATFORM_ID	latitude	longitude	platform_code
	degrees_north	degrees_east	
1049774	53.193787	-7.9936466	
1051983	54.487747	-8.101534	
1051984	51.89995	-8.66196	
1051985	53.360546	-6.490446	
1051986	52.761547	-8.475845	
1054108	44.8883	11.60807	
1057510	38.66022	-0.098842	
1057511	39.88354	-0.514598	
1057512	39.18825	-0.40995	
1057513	39.959667	-0.122058	
1057514	39.52132	-0.505554	
1057515	38.87065	-0.283408	
1064733	37.3453	-2.172287	
1064734	36.72691	-4.615234	

Figure 8 ERDDAP list of platforms in EMODnet Physics

If the user wants to see the platform page from the mapviewer (dev), he/she can use the following construct:

https://map.emodnet-physics.eu/platformpage/?platformid=PLATFORM_ID

3.3.1.4.3 Widgets API

Widgets are accessible from the map viewer through the Platform page. The general widget construct is:
https://map.emodnet-physics.eu/plotWidget/?platformcode=PLATFORMCODE&param=PARA&days=N

where

- PLATFORMCODE is the EMODnet Physics id,
- PARA is the parameter
  - TEMP = temperature
  - PSAL = salinity
  - SLEV = sea level
  - WIND = wind
  - WAVE = wave
  - RVFL = river outflow
  - Other parameters have to be activated yet.
- N is the number of days

3.3.1.4.4 GeoServer

The second core data interoperability technology in EMODnet Physics next to ERDDAP is GeoServer that implements the OGC WMS, WFS, WCS and WMTS standard services:

https://geoserver.emodnet-physics.eu

The available WMS features are:

GetCapabilities	Available for all gridded layers	retrieves metadata about the service, including supported operations and parameters, and a list of the available layers
GetMap	Available for all layers	returns a map image containing icons or text description. returns a subset of data by means of an image depending on the scale/extension.
GetFeatureInfo	Enabled for each layer and timestamp.	□ offer XML, JSON and plain text formats as output. □ returns the stations properties. □ returns measurements.

GetCapabilities

https://prod-geoserver.emodnet-physics.eu/geoserver/ows?service=wms&version=1.3.0&request=GetCapabilities

GetCapabilities – Gridded layer - EMODnet Physics ERDDAP/ncWMS

Construct:

https://ERDDAP-ENDPOINT /ncWMS2/wms?SERVICE=WMS&REQUEST=GetCapabilities&VERSION=1.3.0&DATASET= DATASETNAME

Where:

- ERDDAP-ENDPOINT can be:
  - erddap.emodnet-physics.eu
  - prod-erddap.emodnet-physics.eu/erddap
- DASETNAME is one of the available datasets (see Section **Error! Reference source not found.**)

Example:

https://erddap.emodnet-physics.eu/ncWMS2/wms?SERVICE=WMS&REQUEST=GetCapabilities&VERSION=1.3.0&DATASET=EMODNET_SEA_LEVEL_MONTHLY_MEAN

GetFeatureInfo – Gridded layer - EMODnet Physics ERDDAP/ncWMS

Construct:

https://ERDDAP-ENDPOINT /ncWMS2/wms?REQUEST=GetFeatureInfo&VERSION=1.3.0&STYLES=&CRS=CRS:84&WIDTH=256&HEIGHT=256&I=128&J=128&INFO_FORMAT=text/xml&QUERY_LAYERS= DASETNAME&LAYERS= DASETNAME &BBOX=-180.0,-90.0, 180.0,90.0

Where:

- ERDDAP-ENDPOINT can be:
  - erddap.emodnet-physics.eu
  - prod-erddap.emodnet-physics.eu/erddap
- DASETNAME is one of the available datasets (see Section **Error! Reference source not found.**)

The user can also specify a custom BBOX.

Example:

https://erddap.emodnet-physics.eu/ncWMS2/wms?REQUEST=GetFeatureInfo&VERSION=1.3.0&STYLES=&CRS=CRS:84&WIDTH=256&HEIGHT=256&I=128&J=128&INFO_FORMAT=text/xml&QUERY_LAYERS=EMODNET_SEA_LEVEL_MONTHLY_MEAN/sla&LAYERS=EMODNET_SEA_LEVEL_MONTHLY_MEAN/sla&BBOX=-180.0,-90.0,180.0,90.0

3.3.1.4.5 EMODnet Physics GeoNetwork

A GeoNetwork implementation of EMODnet catalogue is also available. GeoNetwork implements WxS, OGC, ISO standards.

https://prod-catalogue.emodnet-physics.eu

3.3.1.5 Any new developments underway

As part of optimising the exchange from EMODnet Physics services to the EMODnet Central Portal, a re-organisation is underway for managing and publishing the various platforms and their data streams. This

re-organisation is also taking into account making the EMODnet Physics services more fit for purpose of the Blue-Cloud DD&AS conceptual approach. This will be detailed in the later paragraph on 'analysis'.

3.3.2 Standards in use

Many challenges (still) exist associated with good management of ocean data, such as using different formats, wide diversity of datasets, and disparate data management structures, among others (Tanhua *et al.*, 2019). Starting from the Global Ocean Data Assimilation Experiment (GODAE) (Le Traon et al., 2001) many steps have been achieved and the main challenge today is to ensure long term sustainability and add new components to the system for managing emerging network sensors. Anyhow the basic milestones are set and can be easily adopted and adapted to the purpose. This comprises a data processing, validation and dissemination infrastructure. This infrastructure has not to be considered as a monolithic system but it has to implement the latest data federation and interoperability methodologies and data processing centers (DAC) or thematic assembly centers (TAC) are an essential component of this global operational oceanography infrastructure. Data centres have been developed nationally, regionally and globally for many decades, not only as data repositories but also as service providers (Bourtzis, 2015). According to the monitoring program (and monitored ocean essential variables) data may end up into the National Oceanographic Data Center (NODC), and/or into the Thematic Data Assembly Center and/or into a Global Data Assembly Center (GDAC). These centers maintain, update, and provide access to marine environmental and ecosystem data and information. These hubs of data are applying harmonization, nevertheless they are still making metadata and data available according different data transport formats, different standards, and different data publication services. As general examples: CMEMS-INSTAC is offering data via the CMEMS Dissemination Unit (CMEMS DU) and more specifically it implements a dedicated FTP folder for IN SITU data; Coriolis is making data available on a public IFREMER ftp folder; IOC offers html page with latest 30 days of sea level data from IOC-Tide Gauges, etc. EMODnet Physics is an integrator and it connects these hubs applying a further level of metadata harmonization and when a new source is identified, whenever possible, the action is to streamline the new data flow towards one of the major near real time pathways that EMODnet Physics uses to connect sources, and in general, there are three near real time (NRT) pathways to EMODnet Physics. At European level, EuroGOOS that has established Task Teams, and TT members collaborate in the areas of shared priorities, exchange best practices, and feed data to the EuroGOOS ROOS regional portals, and European Marine data programs and initiatives. Most of these platforms already fall into or can be associated to one of the GOOS platforms networks (Data Buoy Cooperation Panel (DBCP)[5], Global Ocean Ship-Based Hydrographic Investigations

[5] *https://www.ocean-ops.org/dbcp/*

Programme (GO-SHIP)[6], Ship Observations Team (SOT)[7], ARGO[8], OceanSITES[9], OceanGliders[10], Animal-Borne Ocean Sensors (AniBOS)) that organize the standards and best practices for the platform stakeholders. The ocean data management and exchange process within EuroGOOS are intended to reduce duplication of effort among agencies, improve quality and reduce costs related to geographic information, thus making oceanographic data more accessible to the public and helping to establish key partnerships to increase data availability. EMODnet Physics adopts and adapts EuroGOOS task teams recommendations that have been lately updated by the H2020 AtlantOS[11] project and H2020 EuroSEA project.

3.3.2.1 Metadata format(s) - short overview and references to detailed documentation

Metadata require the use of standardised vocabularies to solve ambiguities problems. EMODnet Physics uses metadata to describe the dataset and this goes back-to-back to the dataset itself (global attributes of the transport format or metadata in the ERDDAP dataset). For DM data EMODnet Physics relays on the SeaDataNet CDI service and its SeaDataNet conventions.

The metadata format can be checked in the ERDDAP service of EMODnet Physics.

3.3.2.2 Data format(s) - short overview and references to detailed documentation

Together with the metadata comes the data model. Major European marine data infrastructures for operational oceanography are adopting/adapting the OceanSITES data model (with some extensions to provide the user with more information). Typically, the transport format for operational data is NetCDF (CF Convention). This convention includes hierarchical levels of metadata that gives information about who, when, etc (global attributes) and what, how, etc. (variables, sensors attributes, data quality, etc.).

3.3.2.3 Use of controlled vocabularies - which, where, how

Table 6 Controlled Vocabularies as used by EMODnet Physics

Metadata field	Vocabulary exists	Link to vocabulary	Vocabulary governance
Platform id		https://www.ocean-ops.org/	OCEANOPS/WMO

[6] *https://www.go-ship.org/*

[7] *https://www.ocean-ops.org/sot/*

[8] *https://argo.ucsd.edu/*

[9] *http://www.oceansites.org/*

[10] *https://www.oceangliders.org/*

[11] *https://archimer.ifremer.fr/doc/00370/48139/48242.pdf*

Metadata field	Vocabulary exists	Link to vocabulary	Vocabulary governance
		https://vocab.ices.dk/?ref=1399 https://eurogoos.eu/download/eu-hfradar-inventory-2016/?wpdmdl=9972&refresh=642bf4a58042f1680602277 https://www.ego-network.org/dokuwiki/doku.php?id=public:glidersdeployments https://www.ferrybox.org/routes_data/routes/table_of_routes/index.php.en http://eutgn.marine.ie/geonetwork/srv/ita/catalog.search#/home	ICES, EU HFR node EGO (glider) FB Tide Gauge
naming_authority	Yes	https://edmo.seadatanet.org/	SeaDataNet
Institution	Yes	https://edmo.seadatanet.org/	SeaDataNet
qc_method	*	doi	
data_mode	Yes	NRT/DM/REP	EuroGOOS DATAMEQ
variable names	Yes	http://vocab.nerc.ac.uk/collection/P02/current/ http://vocab.nerc.ac.uk/collection/P01/current/ http://vocab.nerc.ac.uk/collection/P07/current/ https://cfconventions.org /Data/cf-standard-names/79/build/cf-standard-name-table.html	BODC:NVS CF Standard Name Table v29
unit	yes	https://vocab.nerc.ac.uk/collection/P06/current/	SeaDataNet
Quality Flag Scheme	yes	http://www.oceansites.org/docs/oceansites_data_format_reference_manual.pdf https://vocab.seadatanet.org/v_bodc_vocab_v2/search.asp?lib=L20	OceanSites SeaDataNet
Time	yes	ISO8601	ISO
Datum	Yes	WGS84	ISO
Country	yes	ISO3166	ISO
Licence	Yes	https://creativecommons.org/	CC

Metadata field	Vocabulary exists	Link to vocabulary	Vocabulary governance
INSPIRE	Yes	ISO 19115	ISO/INSPIRE
PI	yes	https://orcid.org/	ORCID

3.3.2.4 Data access policy - if yes, which and how deployed, using any AAI

All the data sets as published through EMODnet Physics are open and have a license as Unrestricted or CC-BY.

3.3.3 Hosting environment

EMODnet Physics addresses the challenge of ensuring open access to harmonized integrated data across a very diverse and large range of ocean observing "systems with multi-scale and multi-platform/sensor observations. EMODnet Physics develops a common procedure to access ocean physics parameters and products from several sources: available products are collections of in situ data, reanalysis and trends from in situ data, elaboration in space and/or time of in situ data and model output for a given parameter. Data products delivery mode ranges from real-time, near real time to validated long term time series.

To this end, the EMODnet Physics data infrastructure is now logically divided in three layers: 1) data layer, 2) application layer and 3) service layer. The data layer is in charge for linking and harvesting datasets from the various data sources (which can make available data in different formats – csv, txt, netcdf etc - and by means of different distribution tools – FTP, http download, webREST, etc).

EMODnet Physics applies routines to harmonise data and metadata - it applies common vocabularies to complete metadata, applies a common data transport format, etc - and makes data ready for the next layer.

A single dataset-source may be integrated in more products and the service layer organizes data within the EMODnet Physics datasets and data products publication services and catalogues (ERDDAP, GeoServer, GeoNetwork).

The current EMODnet Physics backend infrastructure includes:

- VM1: (2 Core - 8 GB RAM) - 1 docker - runs the endpoint for the ERDDAP server and Apache (it is the external proxy)
- VM2: (4 Core - 16 GB RAM) - 3 dockers – 3 ERDDAP server
- VM3: (12 GB RAM) – 1 docker - internal ERDDAP server (for source linkage and, whenever needed, data buffering)
- VM4: (6 Core - 28 GB) – 1 docker - Python scripts for metadata checks/data format conversion (unification)

- VM5: (1 Core - 4 GB) – 1 docker - microservice queues manager (RabbiMQ)
- VM6: (4 Core - 8 GB) – 1 docker - internal microservices jobs (vocabulary checker, metadata checker, source checker)
- VM7: (2 Core - 8 GB) – 1 docker - internal monitoring (accessibility/performance/service avaialbility) logs
- VM9: (2 Core - 8 GB) – docker – Geoserver
- VM8: (2 Core - 8 GB) – PostgresSQL – geoDB for some of the EMODnet Physics layers
- SAAS MongoDB - (to take over VM8) and host EMODnet Physics metadata

3.3.4 Organisational aspects

3.3.4.1 Main operator(s); data providers

EMODnet Physics relies on an active collaboration with many data providers and data integrators. A major cooperation is with **EuroGOOS** and with the **Copernicus Marine Service INS TAC** data. Under these frameworks data streams from several hundreds of real time stations with multiple parameters are collected and published through EMODnet Physics. While for delayed mode data, there is a structured collaboration with **SeaDataNet**.

In addition, EMODnet Physics federates and publishes data streams from the following national, regional, European, and international observing systems.

Table 7 Overview of additional observing networks feeding EMODnet Physics

Description	Data provider	Tipe of platform/data
National tide gauge network	ISPRA	Tidegauge
Tzunami Array – Devices	ISPRA/JRS	Tidegauge
Rijkswaterstaat observation networks	Rijkswaterstaat	Various
Ice Profiler instruments	Ocean Network Canada	Various
POSEIDON observing network	HCMR	Various
NODB	NOC-BODC	Various
Global Tropical Moored Buoy Array Program	NOAA; Ocean Sites	Various
INCOIS	INCOIS	Moored buoy
CORIOLIS	Ifremer	Various

Description	Data provider	Tipe of platform/data
NOAA's Atlantic Oceanographic and Meteorological Laboratory (AOML)	NOAA	Drifting Buoys
Voluntary Observing Ship Climate (VOSClim) Fleet observations	NOAA	Ship data
Global Sea Level Observing System (GLOSS)	GLOSS	Mooring time series
Coastal Data Information Program (CDIP)	Scripps Institution of Oceanography	Mooring time series
OceanSITES	NOAA NDBC	Mooring time series
European Multidisciplinary Seafloor and water column Observatory (EMSO).	EMSO	Sea flor platform
Marine Weather network	ISMAR	Meteo Stations
Voluntary Observing Ship (VOS) project.	NOAA	Ship data
multi-agency U.S. Animal Telemetry Network (ATN)	NOAA	Animals
Global Runoff Data Centre	Copernicus/GCOS	
MOOSE	Coriolis	Various
TIdal CONstants (TICON) data set	Various	Various
OBSEA	Universitat Politecnica de Catalunya	underwater observatory
EuroGOOS Tide Gauge Task Team – EuroSea tide gauge metadata catalogue for all permanent tide gauges along European and adjacent coastlines	EuroGOOS Various	Tide Gauge
CALYPSO South	Interreg - Various	Models
Saildrone uncrewed surface vehicles (USVs)	NOAA / Saildrone	Saildrone

3.3.4.2 Contact details for technical developments

EMODnet Phyics runs a central focal point for technical developments:
helpdesk@emodnet-physics.eu

3.3.5 Your suggestions for aggregating data sets as collections in order to fit the two-step approach of the Blue-Cloud DD&AS, going from collections to granules

As earlier indicated, EMODnet Physics is working on a re-organisation of its data management structure and the way it is giving access and publishing all the data streams and data products that it is gathering,

federating, and processing. This re-set is required to optimise the exchange from EMODnet Physics services to the new EMODnet Central Portal. Moreover, a win-win situation can be achieved, as the new set-up will also be made fit-for-purpose of the Blue-Cloud DD&AS federation with its two level approach. The following gives the plan of re-organisation that EMODnet Physics currrently is deploying and which might be fully operational by end 2023.

EMODnet Physics is considering the following definitions:

1) data is a series of values (e.g., timeseries, profiles) sampled by an in-situ platform,
2) data collection is a grouping of similar in situ data (e.g., all CTD data, all sea level data with a sampling frequency of at least one value every 5 minutes, sea surface currents from HFR, in-situ data from a specific source),
3) product is the outcome of a reprocessing method, such as the PSMSL (monthly means) RLR or the Coriolis Ocean dataset for Reanalysis (CORA). The outcome of a numerical model (that uses in situ data) is a product. The result of QC/QF procedure is not considered a product; instead, it is a qualified dataset or qualified data collection.

These are the identified package to be served to the Blue Cloud infrastructure.

To move towards a better organization of the data flow and facilitate the process to become part of the federated Blue-Cloud Data Discovery & Access Service, the EMODnet Physics data collection management and publication is being improved. More specifically, the ERDDAP (primary Data discovery and access service component for EMODnet Physics) is now offering two endpoints:

- one for offering products
  https://prod-erddap.emodnet-physics.eu/erddap/index.html
- and one for offering the data collections
  https://data-erddap.emodnet-physics.eu/erddap

The latter endpoint is offering the data collections and publishing a 'metadata' dataset to provide the user with information such as the list of platforms available in the datasets, the providers of those platforms, and the first available measurement, etc. As before, use will be made of common standards and the vocabularies as described earlier in $3.3.2.3. Also, a mapping has been made for each parameter for each of the vocabularies.

According to this convention, the metadata dataset will provide Blue Cloud with L1 information at collection level, while the associated data sets will provide Blue Cloud with the L2 information and access to the data sets at granular level.

An example is given below for river data (RVFL):

https://data-erddap.emodnet-physics.eu/erddap/search/index.html?page=1&itemsPerPage=1000&searchFor=RVFL

EMODnet Physics lists two collections of datasets:

Grid DAP Data	Sub-set	Table DAP Data	Make A Graph	WMS	Source Data Files	Title	Sum-mary	FGDC, ISO, Metadata	Back-ground Info	RSS	E mail	Institution	Dataset ID
	set	data	graph			EMODnet Physics - Collection of river flow rate (RVFL) TimeSeries - MultiPointTimeSeriesObservation		F I M	background	RSS		EMODnet Physics	ERD_EP_TS_RVFL_NRT
	set	data	graph		files	EMODnet Physics - Collection of river flow rate (RVFL) TimeSeries - MultiPointTimeSeriesObservation - METADATA		F I M	background	RSS		EMODnet Physics	ERD_EP_TS_RVFL_NRT_METADATA

The information in the table above is also available in other file formats (.csv, .htmlTable, .itx, .json, .jsonlCSV1, .jsonlCSV, .jsonlKVP, .mat, .nc, .nccsv, .tsv, .xhtml) via a RESTful web service.

Figure 9 Example of ERDDAP results at collection (L1) level

The following table provides the metadata format (e.g. ERD_EP_TS_RVFL_NRT_METADATA) describing the what, when, how, who, of each collection:

Table 8 Overview of ERDDAP attributes describing a collection of EMODnet Physics

Row Type	Variable Name	Attribute Name	Data Type	Value
attribute	NC_GLOBAL	cdm_data_type	String	Other
attribute	NC_GLOBAL	Conventions	String	COARDS, CF-1.10, ACDD-1.3, NCCSV-1.2
attribute	NC_GLOBAL	Easternmost_Easting	double	27.2607
attribute	NC_GLOBAL	geospatial_lat_max	double	70.13984680175781
attribute	NC_GLOBAL	geospatial_lat_min	double	-31.651477813720703
attribute	NC_GLOBAL	geospatial_lat_units	String	degrees_north
attribute	NC_GLOBAL	geospatial_lon_max	double	27.2607
attribute	NC_GLOBAL	geospatial_lon_min	double	-123.18345642089844
attribute	NC_GLOBAL	geospatial_lon_units	String	degrees_east
attribute	NC_GLOBAL	infoUrl	String	https://emodnet.ec.europa.eu/en/physics
attribute	NC_GLOBAL	institution	String	EMODnet Physics
attribute	NC_GLOBAL	keywords	String	assembly, available, center, cod, code, codes, contact, coordinates, country, coverage, creation, CreationDate, data, DataAssemblyCenter, DataFeatureType, DataOwnerCountryCod, DataOwnerCountryName, DataOwnerEdmo, DataOwnerLogo, DataOwnerName, DataOwnerWebSite, date, dates, depth, depths, description, documentation, doi, edmo, end, feature, first, geo, geometry, GeometryCoordinates, GeometryNames,

Row Type	Variable Name	Attribute Name	Data Type	Value
				GeometryType, GeoSpatialAvailableDates, GeoSpatialAvailableDepthCodes, GeoSpatialAvailableDepths, GeoSpatialLastDateObservation, GeoSpatialLastLatitudeObservation, GeoSpatialLastLongitudeObservation, GeoSpatialLatitudeObservation, GeoSpatialLatMax, GeoSpatialLatMin, GeoSpatialLongitudeObservation, GeoSpatialLonMax, GeoSpatialLonMin, GeoSpatialTimeCoverageEnd, GeoSpatialTimeCoverageStart, GeoSpatialVerticalMax, GeoSpatialVerticalMin, group, groups, hash, info, institution, InstitutionCountryCod, InstitutionCountryName, InstitutionEdmo, InstitutionLink, InstitutionName, LastDateObservation, LastDateObservationDT, latitude, link, local, logo, longitude, max, meteorological, min, name, names, observation, organisation, owner, p01, p02, page, parameter, parameters, ParametersInfoGroupP02, ParametersInfoP01, ParametersInfoParameterGroups, ParametersInfoParameters, pi-name, PI_Name, platform, PlatformCode, PlatformCodeHash, PlatformName, PlatformPage, PlatformTypeCode, PlatformTypeDescription, PlatformTypeSDNL06, projects, range, related, sdnl06, site, source, spatial, start, stato, time, TimeRange, type, update, UpdateDate, vertical, web, wmo, world
attribute	NC_GLOBAL	license	String	CC-BY
attribute	NC_GLOBAL	Northernmost_Northing	double	70.13984680175781
attribute	NC_GLOBAL	sourceUrl	String	(local files)
attribute	NC_GLOBAL	Southernmost_Northing	double	-31.651477813720703
attribute	NC_GLOBAL	standard_name_vocabulary	String	CF Standard Name Table v70
attribute	NC_GLOBAL	subsetVariables	String	PLATFORMCODE
attribute	NC_GLOBAL	summary	String	EMODnet Physics - Collection of river flow rate (RVFL) TimeSeries - MultiPointTimeSeriesObservation - METADATA
attribute	NC_GLOBAL	title	String	EMODnet Physics - Collection of river flow rate (RVFL) TimeSeries - MultiPointTimeSeriesObservation - METADATA
attribute	NC_GLOBAL	Westernmost_Easting	double	-123.18345642089844

Row Type	Variable Name	Attribute Name	Data Type	Value
variable	PLATFORMCODE		String	
attribute	PLATFORMCODE	long_name	String	EMODnet Platform Code
variable	call_name		String	
attribute	call_name	long_name	String	Platform Call Name
variable	latitude		double	
attribute	latitude	_CoordinateAxisType	String	Lat
attribute	latitude	actual_range	double	-31.651477813720703, 70.13984680175781
attribute	latitude	axis	String	Y
attribute	latitude	ioos_category	String	Location
attribute	latitude	long_name	String	Latitude
attribute	latitude	standard_name	String	latitude
attribute	latitude	units	String	degrees_north
variable	longitude		double	
attribute	longitude	_CoordinateAxisType	String	Lon
attribute	longitude	actual_range	double	-123.18345642089844, 27.2607
attribute	longitude	axis	String	X
attribute	longitude	ioos_category	String	Location
attribute	longitude	long_name	String	Longitude
attribute	longitude	standard_name	String	longitude
attribute	longitude	units	String	degrees_east
variable	fristDateObservation		double	
attribute	fristDateObservation	ioos_category	String	Time
attribute	fristDateObservation	long_name	String	first Date Observation
attribute	fristDateObservation	standard_name	String	time
attribute	fristDateObservation	time_origin	String	01-JAN-1970 00:00:00
attribute	fristDateObservation	time_precision	String	1970-01-01T00:00:00Z
attribute	fristDateObservation	units	String	seconds since 1970-01-01T00:00:00Z
variable	lastDateObservation		double	
attribute	lastDateObservation	ioos_category	String	Time
attribute	lastDateObservation	long_name	String	last Date Observation
attribute	lastDateObservation	standard_name	String	time
attribute	lastDateObservation	time_origin	String	01-JAN-1970 00:00:00
attribute	lastDateObservation	time_precision	String	1970-01-01T00:00:00Z
attribute	lastDateObservation	units	String	seconds since 1970-01-01T00:00:00Z
variable	parameters_group_longname		String	

Row Type	Variable Name	Attribute Name	Data Type	Value
attribute	parameters_group_longname	long_name	String	Parameters Info Parameter Groups
variable	parameters_group_P02		String	
attribute	parameters_group_P02	long_name	String	Parameters Info P02
variable	parameters		String	
attribute	parameters	long_name	String	Parameters Info Parameters
variable	parameters_P01		String	
attribute	parameters_P01	long_name	String	Parameters Info P01
variable	WMO		int	
attribute	WMO	_FillValue	int	2147483647
attribute	WMO	long_name	String	WMO
variable	DOI		String	
attribute	DOI	long_name	String	DOI
variable	data_owner_longname		String	
attribute	data_owner_longname	long_name	String	Data Owner Name
variable	data_owner_country_code		String	
attribute	data_owner_country_code	long_name	String	Data Owner Country Code
variable	data_owner_country_longname		String	
attribute	data_owner_country_longname	long_name	String	Data Owner Country Name
variable	data_owner_EDMO		byte	
attribute	data_owner_EDMO	_FillValue	byte	127
attribute	data_owner_EDMO	actual_range	byte	0, 120
attribute	data_owner_EDMO	long_name	String	Data Owner EDMO Code
variable	data_assembly_center_longname		String	
attribute	data_assembly_center_longname	long_name	String	Data Assembly Center
variable	platform_type_longname		String	
attribute	platform_type_longname	long_name	String	Platform Type
variable	platform_type_SDNL06		String	
attribute	platform_type_SDNL06	long_name	String	Platform Type SDN L06
variable	platformpage_link		String	
attribute	platformpage_link	long_name	String	Platform Page
variable	SOURCE		String	
attribute	SOURCE	long_name	String	Source
variable	integrator_id		String	
attribute	integrator_id	long_name	String	integrator_id

The user can interact with the metadata dataset on the ERDDAP url:

https://data-erddap.emodnet-physics.eu/erddap/tabledap/ERD_EP_TS_RVFL_NRT_METADATA.html

Or directly with a query:

https://data-erddap.emodnet-physics.eu/erddap/tabledap/TS_RVFL_METADATA.htmlTable?PLATFORMCODE%2Ccall_name%2Clatitude%2Clongitude%2CfristDateObservation%2ClastDateObservation%2Cparameters_group_longname%2Cparameters%2Cparameters_P01%2Cdata_owner_EDMO%2Cplatform_type_longname%2Cplatformpage_link

And the system will list the dataset content (i.e. the list of the stations, the parameters, the location, etc), which will serve for the Blue-Cloud level 2 approach.

Table 9 List of stations (Level 2) within a collection of EMODnet Physics

PLATFORMCODE	call_name	latitude	longitude	fristDateObservation	lastDateObservation	parameters_
		degrees_north	degrees_east	UTC	UTC	
AbbevilleSomme	AbbevilleSomme	50.094498	1.8297544	2022-01-10T12:40:00Z	2022-06-02T11:00:00Z	River
AbromollaVegea	AbromollaVegea	56.07419967651367	12.974499702453613	2022-03-16T00:00:00Z	2023-10-08T23:00:00Z	River
AbzacIsle	AbzacIsle	45.02180480957031	-0.12619030475616455	2021-12-28T00:00:00Z	2023-07-11T14:00:00Z	River
ADN-CURRISO	ADN-CURRISO	45.76167	13.49633	2023-02-12T00:00:00Z	2023-08-04T16:41:09Z	Water Temperature,River
AfonTanat-Llanyblodwel	AfonTanat river at Llanyblodwel station (GRDC code 6609538)	52.79436	-3.11022	2023-02-10T09:00:00Z	2023-09-06T11:00:00Z	River
AgdeHerault	AgdeHerault	43.32531	3.4796884	2021-12-22T00:00:00Z	2023-07-19T18:05:00Z	River
Aire-Kildwick	Aire river at Kildwick station (GRDC code 6605440)	53.9073	-1.98464	2023-02-10T09:00:00Z	2023-09-06T11:00:00Z	River
AkerselvaNordmarkvassdraget	AkerselvaNordmarkvassdraget	59.96883010864258	10.787599563598633	2022-03-12T00:00:00Z	2023-10-10T22:00:00Z	River
AlabaleineAlabaleine	AlabaleineAlabaleine	57.88856	-67.60008	2023-01-19T00:00:00Z	2023-08-24T00:00:00Z	River
AllnabadStrathmore	AllnabadStrathmore	58.34717559814453	-4.644620895385742	2018-01-01T00:00:00Z	2023-10-10T22:00:00Z	River
Almen	Almen	52.1633	6.2	2023-02-12T00:00:00Z	2023-10-05T23:50:00Z	River
AlmondbankAlmond	AlmondbankAlmond	56.415276	-3.5135984	2022-06-08T02:30:00Z	2023-10-06T22:45:00Z	River
AlnessAlness	AlnessAlness	57.69601058959961	-4.258760929107666	2018-01-01T00:00:00Z	2023-10-10T23:00:00Z	River
Amay	Amay	50.53	5.31	2023-02-12T00:00:00Z	2023-10-05T23:00:00Z	River
AnctevilleAy	AnctevilleAy	49.10801	-1.4670926	2021-12-22T00:36:00Z	2023-09-22T01:00:00Z	River
AndelGouessant	AndelGouessant	48.484573	-2.568232	2021-12-22T00:00:00Z	2023-08-23T10:05:00Z	River
Angleur	Angleur	50.61	5.61	2023-02-12T00:00:00Z	2023-10-05T23:50:00Z	River
Anker-Polesworth	Anker river at Polesworth station (GRDC code 6606660)	52.62722	-1.61311	2023-02-10T09:00:00Z	2023-09-06T11:00:00Z	River
Anllons-Carballo	Anllons-Carballo	43.21009826660156	-8.692700386047363	2023-07-18T02:00:00Z	2023-10-06T23:50:00Z	River
AnnevilleSaire	AnnevilleSaire	49.635067	-1.2891345	2021-12-22T00:40:00Z	2023-09-22T00:50:00Z	River
AnthiliSpercheios	AnthiliSpercheios	38.856300354003906	22.466999053955078	2022-03-16T00:00:00Z	2023-10-10T23:00:00Z	River,Water Temperature
AntisantiTavignano	AntisantiTavignano	42.18162155151367	9.386916160583496	2021-12-22T00:00:00Z	2023-10-10T13:00:00Z	River
ArboriLiamone	ArboriLiamone	42.11641311645508	8.818493843078613	2021-12-22T00:00:00Z	2023-08-23T12:00:00Z	River
ArbroathBrothock	ArbroathBrothock	56.567230224609375	-2.5881717205047607	2018-01-01T00:00:00Z	2023-10-10T22:30:00Z	River
ArdalMoisani	ArdalMoisani	58.543670654296875	6.497910022735596	2022-03-12T00:00:00Z	2023-10-10T22:00:00Z	River

Notably, for each platform the metadata dataset also includes a link to plots, for example:

https://map.emodnet-physics.eu/platformpage/?platformid=63d8edb1879483001e666aa0

To interact with data (L2), the data collection has to be used and the user can interact with the ERDDAP dataset url:

https://data-erddap.emodnet-physics.eu/erddap/tabledap/ERD_EP_TS_RVFL_NRT.html

which offers the following metadata for the data sets:

Table 10 Overview of ERDDAP attributes describing a data set of EMODnet Physics

Row Type	Variable Name	Attribute Name	Data Type	Value
attribute	NC_GLOBAL	cdm_data_type	String	TimeSeries
attribute	NC_GLOBAL	cdm_timeseries_variables	String	PLATFORMCODE,SOURCE,latitude,longitude
attribute	NC_GLOBAL	citation	String	Data are the property of the producer/owner distributed through EMODnet Physics. EMODnet Physics and the partners are not responsible for improper use
attribute	NC_GLOBAL	contact	String	contacts at emodnet-physics.eu
attribute	NC_GLOBAL	Conventions	String	CF-1.6 OceanSITES-Manual-1.2 Copernicus-InSituTAC-SRD-1.3 Copernicus-InSituTAC-ParametersList-3.0.0, COARDS, ACDD-1.3, NCCSV-1.2
attribute	NC_GLOBAL	creator_type	String	institution
attribute	NC_GLOBAL	Easternmost_Easting	double	27.2607421875
attribute	NC_GLOBAL	ep_parameter_group	String	River
attribute	NC_GLOBAL	featureType	String	TimeSeries
attribute	NC_GLOBAL	geospatial_lat_max	double	61.926998138427734
attribute	NC_GLOBAL	geospatial_lat_min	double	42.585601806640625
attribute	NC_GLOBAL	geospatial_lat_units	String	degrees_north
attribute	NC_GLOBAL	geospatial_lon_max	double	27.2607421875
attribute	NC_GLOBAL	geospatial_lon_min	double	-4.918429851531982
attribute	NC_GLOBAL	geospatial_lon_units	String	degrees_east
attribute	NC_GLOBAL	geospatial_vertical_max	double	12.0
attribute	NC_GLOBAL	geospatial_vertical_min	double	0.0
attribute	NC_GLOBAL	geospatial_vertical_positive	String	down
attribute	NC_GLOBAL	geospatial_vertical_units	String	m

Row Type	Variable Name	Attribute Name	Data Type	Value
attribute	NC_GLOBAL	infoUrl	String	http://www.emodnet-physics.eu
attribute	NC_GLOBAL	institution	String	EMODnet Physics
attribute	NC_GLOBAL	keywords_vocabulary	String	GCMD Science Keywords
attribute	NC_GLOBAL	license	String	Creative Commons Attribution Share-Alike http://www.opendefinition.org/licenses/cc-by-sa
attribute	NC_GLOBAL	metadata_dataset	String	https://er2webapps.emodnet-physics.eu/erddap/tabledap/TS_RVFL_COPERNICUS_METADATA.html
attribute	NC_GLOBAL	metadata_document	String	https://metadata.emodnet-physics.eu/datasetfile/emodnet/TS_RVFL_INSTAC
attribute	NC_GLOBAL	Northernmost_Northing	double	61.926998138427734
attribute	NC_GLOBAL	publisher_type	String	institution
attribute	NC_GLOBAL	qc_reference_table	String	https://er2webapps.emodnet-physics.eu/erddap/tabledap/qc_reference_table.html
attribute	NC_GLOBAL	SDN	String	SDN:P01::RFDSCH01
attribute	NC_GLOBAL	sourceUrl	String	(local files)
attribute	NC_GLOBAL	Southernmost_Northing	double	42.585601806640625
attribute	NC_GLOBAL	standard_name_vocabulary	String	CF Standard Name Table v70
attribute	NC_GLOBAL	stato	String	A
attribute	NC_GLOBAL	subsetVariables	String	PLATFORMCODE,SOURCE
attribute	NC_GLOBAL	summary	String	EMODnet Physics - Collection of river flow rate (RVFL) TimeSeries - MultiPointTimeSeriesObservation
attribute	NC_GLOBAL	testOutOfDate	String	now-7days
attribute	NC_GLOBAL	time_coverage_end	String	2023-10-07T05:00:00Z
attribute	NC_GLOBAL	time_coverage_start	String	2023-01-14T06:00:00Z

Row Type	Variable Name	Attribute Name	Data Type	Value
attribute	NC_GLOBAL	title	String	EMODnet Physics - Collection of river flow rate (RVFL) TimeSeries - MultiPointTimeSeriesObservation
attribute	NC_GLOBAL	Westernmost_Easting	double	-4.918429851531982
variable	PLATFORMCODE		String	
attribute	PLATFORMCODE	cf_role	String	timeseries_id
attribute	PLATFORMCODE	long_name	String	EMODnet Platform Code
variable	SOURCE		String	
attribute	SOURCE	long_name	String	source
variable	SENSOR		String	
attribute	SENSOR	long_name	String	Platform Sensor
variable	time		double	
attribute	time	_CoordinateAxisType	String	Time
attribute	time	actual_range	double	1.673676E9, 1.6966548E9
attribute	time	axis	String	T
attribute	time	calendar	String	Gregorian
attribute	time	ioos_category	String	Time
attribute	time	long_name	String	Valid Time GMT
attribute	time	standard_name	String	time
attribute	time	time_origin	String	01-JAN-1970 00:00:00
attribute	time	units	String	seconds since 1970-01-01T00:00:00Z
variable	TIME_QC		short	
attribute	TIME_QC	_FillValue	short	-32767

Row Type	Variable Name	Attribute Name	Data Type	Value
attribute	TIME_QC	actual_range	short	1, 1
attribute	TIME_QC	flag_meanings	String	no_qc_performed good_data probably_good_data bad_data_that_are_potentially_correctable bad_data value_changed not_used nominal_value interpolated_value missing_value
attribute	TIME_QC	long_name	String	TIME quality flag
attribute	TIME_QC	units	String	1
attribute	TIME_QC	valid_range	short	0, 9
variable	depth		double	
attribute	depth	_CoordinateAxisType	String	Height
attribute	depth	_CoordinateZisPositive	String	down
attribute	depth	_FillValue	double	9.969209968386869E36
attribute	depth	actual_range	double	0.0, 12.0
attribute	depth	axis	String	Z
attribute	depth	ioos_category	String	Location
attribute	depth	long_name	String	Depth
attribute	depth	positive	String	down
attribute	depth	standard_name	String	depth
attribute	depth	units	String	m
variable	DEPTH_QC		short	
attribute	DEPTH_QC	_FillValue	short	-32767
attribute	DEPTH_QC	actual_range	short	-127, 9
attribute	DEPTH_QC	flag_meanings	String	no_qc_performed good_data probably_good_data bad_data_that_are_potentially_correctable bad_data value_changed not_used nominal_value interpolated_value missing_value

Funded by the European Union
Blue-Cloud2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227

Row Type	Variable Name	Attribute Name	Data Type	Value
attribute	DEPTH_QC	long_name	String	DEPTH quality flag
attribute	DEPTH_QC	standard_name	String	depth
attribute	DEPTH_QC	units	String	1
attribute	DEPTH_QC	valid_range	short	0, 9
variable	latitude		double	
attribute	latitude	_CoordinateAxisType	String	Lat
attribute	latitude	_FillValue	double	9.969209968386869E36
attribute	latitude	actual_range	double	42.585601806640625, 61.926998138427734
attribute	latitude	axis	String	Y
attribute	latitude	ioos_category	String	Location
attribute	latitude	latitude_reference_datum	String	geographical coordinates, WGS84 projection
attribute	latitude	long_name	String	Latitude
attribute	latitude	standard_name	String	latitude
attribute	latitude	units	String	degrees_north
attribute	latitude	valid_max	double	90.0
attribute	latitude	valid_min	double	-90.0
variable	longitude		double	
attribute	longitude	_CoordinateAxisType	String	Lon
attribute	longitude	_FillValue	double	9.969209968386869E36
attribute	longitude	actual_range	double	-4.918429851531982, 27.2607421875

Row Type	Variable Name	Attribute Name	Data Type	Value
attribute	longitude	axis	String	X
attribute	longitude	ioos_category	String	Location
attribute	longitude	latitude_reference_datum	String	geographical coordinates, WGS84 projection
attribute	longitude	long_name	String	Longitude
attribute	longitude	standard_name	String	longitude
attribute	longitude	units	String	degrees_east
attribute	longitude	valid_max	double	180.0
attribute	longitude	valid_min	double	-180.0
variable	POSITION_QC		short	
attribute	POSITION_QC	_FillValue	short	-32767
attribute	POSITION_QC	actual_range	short	0, 7
attribute	POSITION_QC	flag_meanings	String	no_qc_performed good_data probably_good_data bad_data_that_are_potentially_correctable bad_data value_changed not_used nominal_value interpolated_value missing_value
attribute	POSITION_QC	long_name	String	POSITION quality flag
attribute	POSITION_QC	units	String	1
attribute	POSITION_QC	valid_range	short	0, 9
variable	RVFL		double	
attribute	RVFL	_FillValue	double	9.969209968386869E36
attribute	RVFL	actual_range	double	-2.147483647E9, 1.3720603E7
attribute	RVFL	ep_parameter_group	String	River
attribute	RVFL	long_name	String	river flow rate

Row Type	Variable Name	Attribute Name	Data Type	Value
attribute	RVFL	sample_rate_comment	String	Hourly: 3600, Daily: 8.64e4, Monthly: 2.628e6, Yearly: 3.154e7
attribute	RVFL	sample_rate_units	String	seconds
attribute	RVFL	SDN	String	SDN:P01::RFDSCH01
attribute	RVFL	source_variable_name	String	RVFL
attribute	RVFL	standard_name	String	water_volume_transport_into_sea_water_from_rivers
attribute	RVFL	units	String	m3/s
variable	RVFL_QC		short	
attribute	RVFL_QC	_FillValue	short	-32767
attribute	RVFL_QC	actual_range	short	-127, 4
attribute	RVFL_QC	flag_meanings	String	no_qc_performed good_data probably_good_data bad_data_that_are_potentially_correctable bad_data value_changed not_used nominal_value interpolated_value missing_value
attribute	RVFL_QC	long_name	String	RVFL quality flag
attribute	RVFL_QC	units	String	1
attribute	RVFL_QC	valid_range	short	0, 9
variable	RVFL_DM		char	
attribute	RVFL_DM	actual_range	char	\u0000, R
attribute	RVFL_DM	flag_meanings	String	real-time provisional delayed-mode mixed
attribute	RVFL_DM	flag_values	String	R, P, D, M
attribute	RVFL_DM	long_name	String	RVFL method of data processing
variable	url_metadata		String	
attribute	url_metadata	long_name	String	Metadata Link
variable	qc_entity		int	

Row Type	Variable Name	Attribute Name	Data Type	Value
attribute	qc_entity	actual_range	int	14, 16
attribute	qc_entity	qc_reference_table	String	"https://er2webapps.emodnet-physics.eu/erddap/tabledap/qc_reference_table.htmlTable

Interacting with the page the user can refine the selection and retrieve the full story (e.g. timeseries since day zero) for a given platform. E.g. to have the timeseries for the station “Almen” (the PLATFORMCODE is the primary id linking the metadata dataset and the data collection dataset), one can use the following query:

https://ercompwebapps.emodnet-physics.eu/erddap/tabledap/TS_RVFL.htmlTable?PLATFORMCODE%2Ctime%2CTIME_QC%2Cdepth%2CDEPTH_QC%2Clatitude%2Clongitude%2CPOSITION_QC%2CRVFL%2CRVFL_QC%2CRVFL_DM%2Curl_metadata%2Cqc_entity&PLATFORMCODE=%22Almen%22&time%3E=2023-01-14T06%3A00%3A00Z&time%3C=2023-10-07T05%3A00%3A00Z

This gives the full time series data for the selected station and parameter:

Table 11 Overview of ERDDAP data values in time for one selected station and parameter data set of EMODnet Physics

EMODnet Physics ERDDAP
Easier access to scientific data
EMODnet
log in English
Brought to you by EMODnet Physics

PLATFORMCODE	time	TIME_QC	depth	DEPTH_QC	latitude	longitude	POSITION_QC	RVFL	RVFL_QC	RVFL_DM	
	UTC	1	m	1	degrees_north	degrees_east	1	m3/s	1		
Almen	2023-02-12T00:00:00Z	1		9	52.163299560546875	6.199999809265137	7	5.630000267410651	1	\u0000	https://er2webapps.emodnet-physics.eu/erddap/tabledap/TS_RVFL_IN
Almen	2023-02-12T00:10:00Z	1		9	52.163299560546875	6.199999809265137	7	12.250000581843778	1	\u0000	https://er2webapps.emodnet-physics.eu/erddap/tabledap/TS_RVFL_IN
Almen	2023-02-12T00:20:00Z	1		9	52.163299560546875	6.199999809265137	7	6.370000302558765	1	\u0000	https://er2webapps.emodnet-physics.eu/erddap/tabledap/TS_RVFL_IN
Almen	2023-02-12T00:30:00Z	1		9	52.163299560546875	6.199999809265137	7	2.220000105444342	1	\u0000	https://er2webapps.emodnet-physics.eu/erddap/tabledap/TS_RVFL_IN
Almen	2023-02-12T00:40:00Z	1		9	52.163299560546875	6.199999809265137	7	8.880000421777368	1	\u0000	https://er2webapps.emodnet-physics.eu/erddap/tabledap/TS_RVFL_IN
Almen	2023-02-12T00:50:00Z	1		9	52.163299560546875	6.199999809265137	7	12.620000599417835	1	\u0000	https://er2webapps.emodnet-physics.eu/erddap/tabledap/TS_RVFL_IN
Almen	2023-02-12T01:00:00Z	1		9	52.163299560546875	6.199999809265137	7	7.780000369530171	1	\u0000	https://er2webapps.emodnet-physics.eu/erddap/tabledap/TS_RVFL_IN
Almen	2023-02-12T01:10:00Z	1		9	52.163299560546875	6.199999809265137	7	5.130000243661925	1	\u0000	https://er2webapps.emodnet-physics.eu/erddap/tabledap/TS_RVFL_IN
Almen	2023-02-12T01:20:00Z	1		9	52.163299560546875	6.199999809265137	7	4.500000213738531	1	\u0000	https://er2webapps.emodnet-physics.eu/erddap/tabledap/TS_RVFL_IN
Almen	2023-02-12T01:30:00Z	1		9	52.163299560546875	6.199999809265137	7	2.190000104019418	1	\u0000	https://er2webapps.emodnet-physics.eu/erddap/tabledap/TS_RVFL_IN
Almen	2023-02-12T01:40:00Z	1		9	52.163299560546875	6.199999809265137	7	7.750000368105248	1	\u0000	https://er2webapps.emodnet-physics.eu/erddap/tabledap/TS_RVFL_IN
Almen	2023-02-12T01:50:00Z	1		9	52.163299560546875	6.199999809265137	7	9.030000428901985	1	\u0000	https://er2webapps.emodnet-physics.eu/erddap/tabledap/TS_RVFL_IN
Almen	2023-02-12T02:00:00Z	1		9	52.163299560546875	6.199999809265137	7	5.6500002683606	1	\u0000	https://er2webapps.emodnet-physics.eu/erddap/tabledap/TS_RVFL_IN
Almen	2023-02-12T02:10:00Z	1		9	52.163299560546875	6.199999809265137	7	3.6700001743156463	1	\u0000	https://er2webapps.emodnet-physics.eu/erddap/tabledap/TS_RVFL_IN
Almen	2023-02-12T02:20:00Z	1		9	52.163299560546875	6.199999809265137	7	7.350000349106267	1	\u0000	https://er2webapps.emodnet-physics.eu/erddap/tabledap/TS_RVFL_IN
Almen	2023-02-12T02:30:00Z	1		9	52.163299560546875	6.199999809265137	7	6.5400003106333315	1	\u0000	https://er2webapps.emodnet-physics.eu/erddap/tabledap/TS_RVFL_IN
Almen	2023-02-12T02:40:00Z	1		9	52.163299560546875	6.199999809265137	7	4.900000232737511	1	\u0000	https://er2webapps.emodnet-physics.eu/erddap/tabledap/TS_RVFL_IN
Almen	2023-02-12T02:50:00Z	1		9	52.163299560546875	6.199999809265137	7	6.750000320607796	1	\u0000	https://er2webapps.emodnet-physics.eu/erddap/tabledap/TS_RVFL_IN
Almen	2023-02-12T03:00:00Z	1		9	52.163299560546875	6.199999809265137	7	6.230000295909122	1	\u0000	https://er2webapps.emodnet-physics.eu/erddap/tabledap/TS_RVFL_IN
Almen	2023-02-12T03:10:00Z	1		9	52.163299560546875	6.199999809265137	7	7.850000372854993	1	\u0000	https://er2webapps.emodnet-physics.eu/erddap/tabledap/TS_RVFL_IN
Almen	2023-02-12T03:20:00Z	1		9	52.163299560546875	6.199999809265137	7	5.480000260286033	1	\u0000	https://er2webapps.emodnet-physics.eu/erddap/tabledap/TS_RVFL_IN
Almen	2023-02-12T03:30:00Z	1		9	52.163299560546875	6.199999809265137	7	4.2000001994892955	1	\u0000	https://er2webapps.emodnet-physics.eu/erddap/tabledap/TS_RVFL_IN
Almen	2023-02-12T03:40:00Z	1		9	52.163299560546875	6.199999809265137	7	6.100000289734453	1	\u0000	https://er2webapps.emodnet-physics.eu/erddap/tabledap/TS_RVFL_IN
Almen	2023-02-12T03:50:00Z	1		9	52.163299560546875	6.199999809265137	7	9.15000043460168	1	\u0000	https://er2webapps.emodnet-physics.eu/erddap/tabledap/TS_RVFL_IN
Almen	2023-02-12T04:00:00Z	1		9	52.163299560546875	6.199999809265137	7	6.730000319657847	1	\u0000	https://er2webapps.emodnet-physics.eu/erddap/tabledap/TS_RVFL_IN
Almen	2023-02-12T04:10:00Z	1		9	52.163299560546875	6.199999809265137	7	6.680000317282975	1	\u0000	https://er2webapps.emodnet-physics.eu/erddap/tabledap/TS_RVFL_IN

A comparable dialogue can be applied to other platforms and parameters.

The overarching Catalogue of sub catalogues can be found at:

https://data-erddap.emodnet-physics.eu/erddap/tabledap/allDatasets.html

Currently, this gives only a limited number of records, as EMODnet Physics is underway with re-organising and restructuring all its data sets and stations to fit the new concept and ERDDAP services.

3.3.6 Any data subsetting services in use or under development – URLs, function, how to operate

The ERDDAP services provide functionality for subsetting. However, considering the large number of data sets and their big volume, EMODnet Physics is also looking into solutions that can support subsetting with high performance and that can work underneath and together with the ERDDAP service for discovery and access. In that framework, EMODnet Physics is exploring the use of the new BEACON tool as being developed by MARIS. Results of this, will be reported in the deliverable **D2.4 -BDI sub-setting APIs and Data Lakes – Concept and Specifications Report** which is planned for end February 2024.

3.3.7 Conclusions for Data Infrastructure

The focus for the Blue-Cloud DD&AS federation will be on the output of the EMODnet Physics ERDDAP services that EMODnet Physics is deploying as part of its re-organisation and which is ongoing for getting all data sets included in the new set-up:

https://data-erddap.emodnet-physics.eu/erddap

This will allow to retrieve the metadata of the EMODnet Physics data collections at Blue-Cloud level 1 and the metadata records and links to data sets for the stations and parameters as related to each data collection for the level 2 granular approach. The linking point from level 1 to level 2 is the PLATFORMCODE for each station which gives access to its metadata at level 2 AND access to the full timeseries of data sets for downloading. Also previews will be provided as plots.

The metadata and data formats as used by EMODnet Physics for level 1 and level 2 are documented (see $3.3.5). When comparing with the Blue-Cloud DAB broker metadata profile (see $2.1), it appears that all required Blue-Cloud metadata fields tags (title, abstract, keywords, bounding box, temporal extent, platform, organisation, parameters, instruments) will be available. This will facilitate the actual mapping that has to be done, both for level 1 and level 2. Moreover, the EMODnet Physics output contains several additional attributes which might be used to expand the Blue-Cloud metadata set.

Considering level 2, further analysis is needed if use should be made of the ERDDAP subsetting ability. A lot of collections are timeseries which are incremental with their master file increasing in time or by adding additional timeperiods in separate data files. It might be that some data sets are getting too big for downloading, although most physical data sets are normally quite lean. A start will be made with the integral full data sets as provided by EMODnet Physics.

Another aspect is semantic interoperability. EMODnet Physics makes use of several controlled vocabularies, such as e.g. from the SeaDataNet - NERC Vocabulary Service (see $3.3.2.3). It needs to be reviewed, if the output already contains not only the literal description, but also the URIs and URNs of vocabulary terms as used. These need to be in place and added by EMODnet Physics, if not already, to anticipate the semantic interoperability requirements that are part of the planned optimisation of the Blue-Cloud DD&AS as documented in Blue-Cloud 2026 deliverable D2.1. This aspect will be further analysed during the mapping exercise for the Blue-Cloud level 1 and level 2.

As said, EMODNet Physics is underway with converting metadata and data from sources to make them more homogeneous and ‘clean’ and also adapting terms to SeadataNet vocabs for variables and EDMO for organisations. This is done at a staging server configuration from which metadata and data is migrated to the ‘production’ environment.

EMODnet Physics has several groups of data sources such as Argo, Gliders, HF-Radar, … and all these groups will come back in the ERDDAP Catalogue of data collections. A number of these groups might be doubling with BDIs that the DD&AS is already federating. Therefore, it should be considered to make use of filtering.

All data distributed by EMODnet Physics are **Open;** by default, and CC-BY is used.

3.4. ELIXIR-MGnify

EBI manages the ELIXIR-ENA data infrastructure which is already part of the federation of the Blue-Cloud Data Discovery & Access Service. The MGnify platform is an analytical component of the ELIXIR-ENA system and also part of the ELIXIR infrastructure.

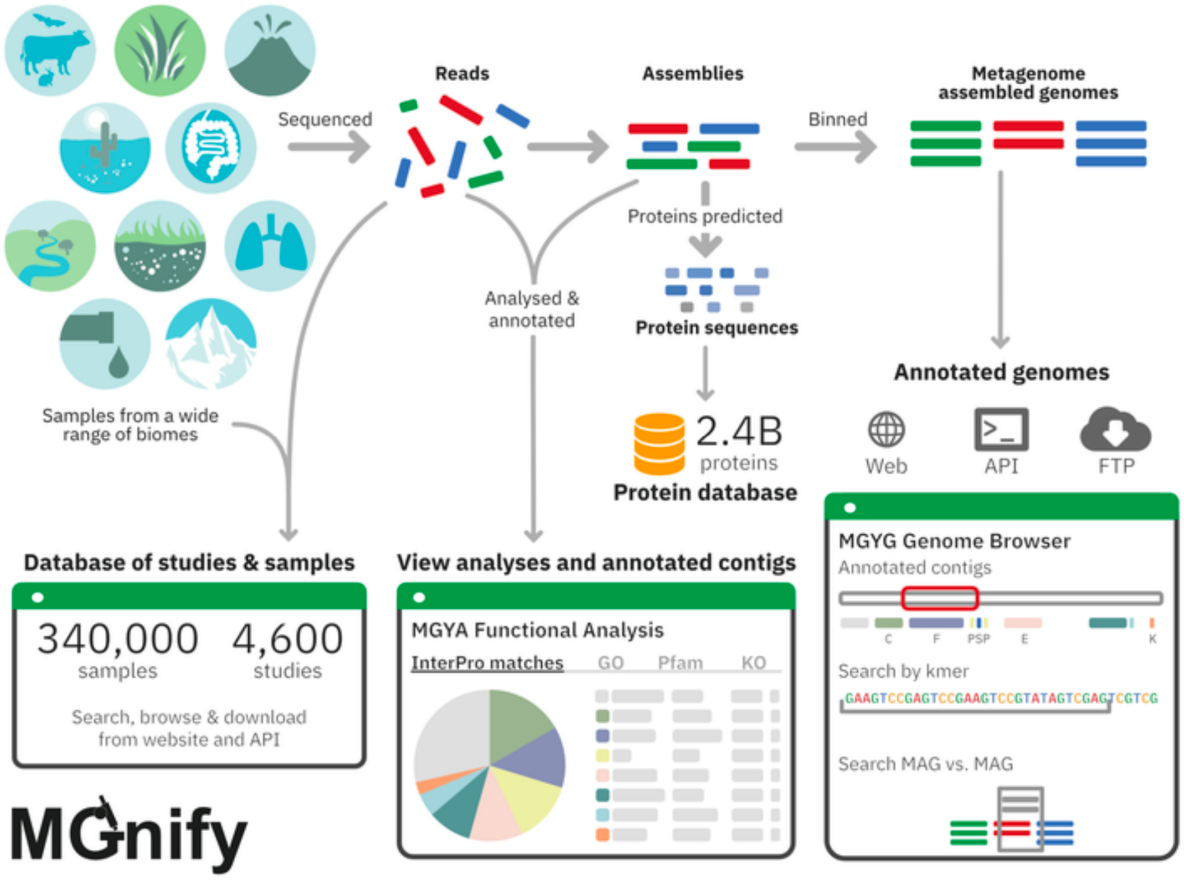


Figure 10 Overview of the ELIXIR-MGnify service

3.4.1. Data discovery and access service component

The MGnify platform facilitates the assembly, analysis and archiving of microbiome-derived nucleic acid sequences. As such the platform provides access to taxonomic assignments and functional annotations for analyses covering metabarcoding, metatranscriptomic, and metagenomic datasets, which are derived from a wide range of different environments. MGnify employs a set of standardised (versioned) analysis pipelines allowing results to be interpreted in context with other datasets. All tools and pipelines are open and freely available within public repositories (https://github.com/EBI-Metagenomics) and all workflows are formally described in Common Workflow Language or Nextflow. These pipelines are deposited in

Funded by the European Union
Blue-Cloud2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227

WorkflowHub (https://workflowhub.eu/projects/9) to support easy reuse within the research community. MGnify works closely with the European Nucleotide Archive (ENA), which archives sample metadata, sequence reads, and assemblies. Researchers can submit pre-publication data to the ENA and request the assembly and/or analysis of those data by MGnify with results subsequently provided within the user's own private area of MGnify. Users may also request the assembly and/or (re-)analysis of any relevant public dataset available in the International Nucleotide Sequence Database Collaboration (INSDC) initiative. To ensure that the resource is fit for purpose, MGnify is a partner in multiple major European Commission funded projects covering a variety of research areas such as Holofood, FindingPheno, AltantECO, BioOcean5D, DTO BioFlow, and BlueRemediomics, and is part of the EMBL-EBI ELIXIR service delivery plan. MGnify is also connected and contributes to other national and international efforts aimed at developing and supporting data standards within the field such as NFDI4 Microbiota, National Microbiome Data Collaborative (NMDC), and Genomic Standards Consortium (GSC).

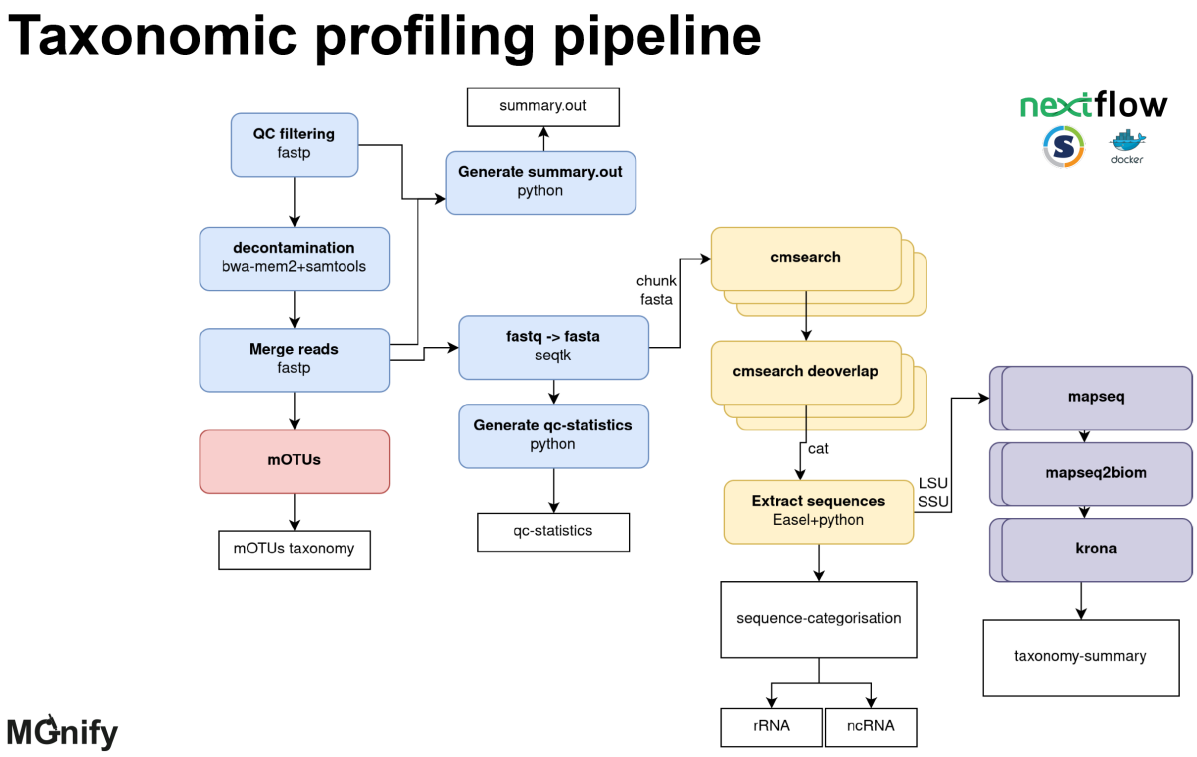


Figure 11 Overview of the MGnify taxonomic profiling pipeline

3.4.1.1 Name and web address

Resource name: MGnify
Web address: https://www.ebi.ac.uk/metagenomics

3.4.1.2 Types and number of data sets and/or data products

MGnify analysis service (version 5.0) offers specialised workflows for three different data types: amplicon, raw metagenomic/metatranscriptomic reads, and assembly.

Table 12 Overview of current number of MGnify data sets/ data products and the marine specific subset

Data/data product type	MGnify total	Marine-specific
Metabarcoding analysis	396307	71750
Metagenomic analysis	37516	4335
Metatranscriptomic analysis	2242	768
Assembled metagenome analysis	35486	4237
Metagenomic Assembled Genomes (MAGs)	315252	1504
Assembled metagenomes	33885	3889
Predicted protein sequence (millions)	2973	777

3.4.1.3 Discovery and access mechanisms - how does it function

The MGnify databases can be accessed via an Application Programming Interface (API) which largely follows the JSON:API specification (https://jsonapi.org). The API is public and can be queried directly by users, and browsed at:

https://www.ebi.ac.uk/metagenomics/api

The MGnify website is a client for the API and allows non-programmatic access via search and browse interfaces:

https://www.ebi.ac.uk/metagenomics

There are several search functions for the MGnify databases:

- EBI Search (a search service across EBI resources) provides a text and facet search across MGnify samples, studies, and analyses, e.g. for finding all studies from marine biomes with a certain phrase in their title
- API endpoint filters provide custom filtering options for lists of data, e.g. for finding samples with depth metadata within a certain range

Funded by the European Union
Blue-Cloud2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227

- Genome search (based on Sourmash) provides a sketch-based comparison of query genomes against MGnify's genome catalogues
- Genome fragment search (based on COBS) provides a kmer search of query genes/sequences against MGnify's genome catalogues
- MGnify Sequence Search provides a HMMER-based sequence search over MGnify's non-redundant protein database

MGnify's Notebooks (https://docs.mgnify.org/src/notebooks_list.html) are an additional client, which provide examples and templates for downstream analysis beyond the features available on the website. For example, the notebooks include interactive examples for cross-study comparative metagenomics.

Some MGnify datasets are also (or only) available via the EBI FTP server (http://ftp.ebi.ac.uk/pub/databases/metagenomics). This provides access to flat-file releases of the protein database, and the full MGnify genome catalogues, including cluster members which are not available via the MGnify API/website.

3.4.1.4 Any new developments underway

EBI is working on a new release of its analysis pipelines. This latest version (version 6.0) will incorporate updates to the existing tools and references databases utilised within the pipeline, as well as providing new functionality. Specifically, within the assembly analysis pipeline, support is added for the prediction and annotation of viral sequences, annotation of the mobilome, and CRISPR-Cas annotations with a view to providing genomic context for the viral annotations. The new version of the metabarcoding analysis pipeline will build on the existing taxonomic classifications by providing amplicon sequence variants (ASV) analysis, of particular interest in the marine and environmental sciences fields, and enables better cross studies comparisons. In recent years EBI has reduced its effort in raw-read analysis due to the much richer annotations that can be provided on assembled contigs. However, in an effort to best support analysis of data from highly diverse environments (that is both time-consuming and computationally expensive to assemble), it is developing a streamlined raw-read pipeline for use alongside the version 6 pipelines. A module has been developed that provides taxonomic profiling and abundance using mOTUs. Combining this module with a minimal functional annotation will simplify, and significantly reduce the computational burden of the existing raw-read analysis pipeline. Further, providing this alongside the rich assembly analysis will result in a comprehensive set of annotations for a given assembled dataset that covers taxonomy, abundance, protein sequence, protein function, pathway annotations, biosynthetic gene clusters, viral and mobilome predictions. Another significant new development underway is version 2.0 of the MGnify Genomes catalogue for the marine biome. In the last year EBI has started to provide biome-specific microbiome-derived genome catalogues as reference sets for metagenomic analysis. Building on work within the AtlantECO project, and BlueRemediomics EBI plans to significantly expand the existing marine catalogue, utilising both MGnify-generated metagenomic-assembled genomes as well as

community-generated genome sets to produce version 2.0 of the marine catalogue. This will cover both prokaryotic and eukaryotic genomes. Finally, in the next year or so, EBI will start to produce viral catalogues.

3.4.2 Standards in use

3.4.2.1 Metadata format(s) - short overview and references to detailed documentation

All metadata presented in MGnify relating the samples and experiments are inherited from the sequence deposition in ENA. Relevant metadata is indexed and both fetchable via the MGnify API, and can be used to filter query results within the webpage filtering and API data retrieval. Additional metadata are incorporated into the website via the EuropePMC metagenomics API which supplies metadata from the literature contained in Europe PMC using a machine learning framework. Metadata pertaining to the analysis pipeline that was run (specific tools and versions etc) are all detailed and linked from the pipeline version description on MGnify (https://www.ebi.ac.uk/metagenomics/pipelines/5.0), as well as via EBI's GitHub repository (https://github.com/EBI-Metagenomics/pipeline-v5).

Research Object Crates (RO-Crates) are in use to store, transfer and display analysis results alongside their provenance metadata. MGnify's RO-Crates follow schemas, e.g

https://www.researchobject.org/workflow-run-crate/

This allows results from workflows beyond MGnify's standardised pipelines to be registered in and retrieved from MGnify. This standard is currently used to store and display biosynthetic gene cluster and mobilome annotations for assemblies alongside MGnify's standard analyses.

3.4.2.2 Data format(s) - short overview and references to detailed documentation

The API endpoints primarily return JSON data.

Many API list endpoints can also return tabular (TSV) data by appending ?format=csv to a URL.

Sequence data are provided in various formats of FASTA file, e.g. nucleotide and protein FASTA, FASTQ. MGnify genomes additionally have FASTA Index (FAI) files.

Annotations from tools in the standard pipelines are typically provided as tabular TSV/CSV files, as well as consolidated into GFF3 files. Some functional annotations are also available in EMBL and Genbank formats.

Taxonomic annotations are available as BIOM files (in both JSON and HDF5 format).

Additional data formats are available on the FTP server for MGnify Genomes, and these are described in each catalogue's README (e.g.

http://ftp.ebi.ac.uk/pub/databases/metagenomics/mgnify_genomes/marine/v1.0/README.txt).

These include Kraken formatted databases for each catalogue.

MGnify's protein database is distributed as FASTA and TSV files (http://ftp.ebi.ac.uk/pub/databases/metagenomics/peptide_database/current_release/

3.4.2.3 Use of controlled vocabularies - which, where, how

Controlled vocabularies and accessioning are supported wherever possible, such as the use of GOLD classification for biome assignment, Gene Ontology terms for functional analysis, and supporting both NCBI and GTDB taxonomic assignments for genome catalogues.

3.4.2.4 Data access policy - if yes, which and how deployed, using any AAI

Most of the data contained in MGnify is open access, under the same licensing model as the ENA (https://rdm.elixir-belgium.org/ena). MGnify's protein database is available under a CC0 1.0 licence.

3.4.3 Web services and API's - URLs, function, how to operate

MGnify is available via the API:

https://www.ebi.ac.uk/metagenomics/api

the website:

https://www.ebi.ac.uk/metagenomics

the FTP server:

http://ftp.ebi.ac.uk/pub/databases/metagenomics/

Documentation is available at:

https://docs.mgnify.org.

3.4.4 Your suggestions for aggregating data sets as collections in order to fit the two-step approach of the Blue-Cloud DD&AS, going from collections to granules

As a new BDI, this is an area where EBI is less knowledgeable. However, EBI would propose making its marine MAG catalogue and analysis results from different samples (taxonomic assertions [ASVs, OTUs, and MAGs mapping to GTDB] , functional profiles [KEGG, GO, InterPro, EggNOG]) available as part of the Blue-Cloud DD&AS offerings. How these connect into the identifier space is unclear, as is the level of granularity required by the user/Blue-Cloud DD&AS.

3.4.5 Any data subsetting services in use or under development – URLs, function, how to operate

The MGnify data model includes Super Studies (similar to umbrella projects in ENA) which can be used to collate a subset of studies related to a certain initiative. For example, all studies from AtlantECO are collated under https://www.ebi.ac.uk/metagenomics/super-studies/atlanteco. Super Studies are manually curated by the MGnify team.

3.4.6 Hosting environment

MGnify's API and website are hosted in EMBL-EBI's Web Production infrastructure. Traffic is load-balanced across multiple compute nodes with primary and fallback data centres. This helps ensure extremely high availability of the resources.

Some secondary services are hosted elsewhere: the COBS-based genome search is hosted on EBI's Embassy Cloud platform; the MGnify Notebooks Server is hosted by EMBL's Bio-IT team using de.NBI infrastructure, and on Galaxy Europe.

3.4.7 Organisational aspects

3.4.7.1 Main operator(s); data providers

MGnify is developed and maintained by a large team of people (currently 13) that are responsible for different aspects of the resource, database, interface and content. The most efficient way to contact to the appropriate operator is to go via the user support helpdesk:

https://www.ebi.ac.uk/about/contact/support/metagenomics

3.4.7.2 Contact details for technical developments

Robert D. Finn (rdf@ebi.ac.uk) - Principle Investigator
Lorna Richardson (lornar@ebi.ac.uk) - MGnify Co-ordinator
Martin Beracochea (mbc@ebi.ac.uk) - MGnify Technical Lead

3.4.8 Conclusions for Data Infrastructure

The ELIXIR-MGnify service is a relatively new development. MGnify contains and manages results of additional analyses on a subset of samples which are described in the ELIXIR-ENA service, which is already federated in the Blue-Cloud DD&AS. Also, MGnify already contains the new extra tags for retrieving the marine, coastal, and inland water related data sets" as described in Blue-Cloud 2026 deliverable D2.1 for the planned optimisation of the ELIXIR-ENA federation in the Blue-Cloud DD&AS. However, MGnify has a separate/different classification of its analyses into biomes, including marine and freshwater. For homogeneity, it is proposed that ENA's environmental partitions (marine, coastal&brackish, freshwater and terrestrial; each with three confidence levels) will be used instead of MGnify's biome classification in order to serve data to the Blue Cloud.

The focus for the Blue-Cloud DD&AS federation will be on the output of the MGnify JSON-API:

https://www.ebi.ac.uk/metagenomics/api

As in the case of ENA, there is not a real separation in metadata and related data sets. In fact, the metadata sheets are the data sets as most results of the sampling and additional analyses for MGnify are qualitative information.

The JSON format of the API service needs to be analysed to get more insight in its structure and the possibilities for a level 1 – level 2 approach. In the case of ENA this is achieved by considering study_accession entries as level 1 collections, while level 2 then are the associated study result records. For MGnify, this needs to be investigated and checked by MARIS and CNR-IIA together with EBI's MGnify team.

Thereby, also focus will be on possible mapping to the Blue-Cloud DAB common metadata profile and beyond, as well as looking into the use of controlled vocabularies, which might have to be more semantically enriched for supporting the semantic brokerage as planned for optimising the Blue-=Cloud DD&AS.

The current API of MGnify already supports subsetting functionality, which is implemented by using Elastic for indexing. Considering the character of the data as included, mostly consisting of metadata, having Elastic is very powerful in subsetting and most appropriate for the type of data.

Most of the data contained in MGnify is open access, under the same licensing model as the ENA (https://rdm.elixir-belgium.org/ena). MGnify's protein database is available under a CC0 1.0 licence.

4. Conclusions and follow-up

The existing Blue-Cloud Data Discovery & Access Service already is federating 9 data discovery and access services from 8 Blue Data Infrastructures (BDIs). As part of the Blue-Cloud 2026 project, activities are underway at central DD&AS level and at the level of each of these BDIs for optimising the overall functionality and services of the DD&AS. The optimisation options as partly already underway and partly further proposed. The latter are documented in the recent deliverable Blue-Cloud 2026 D2.1 – Existing DD&AS and Blue Data Infrastructures – Review and Specifications for Optimisation Report.

Another WP2 task of the Blue-Cloud 2026 project is expanding the DD&AS with services from 4 new BDIs. As part of the trajectory towards such a federation, each of the new infrastructures is described in this deliverable D2.2. The descriptions have a particular focus on their current data discovery and access mechanisms, and how these might fit or made fit for the overall concept of the DD&AS, namely making a distinction, where possible, between collections (level 1) and granular records including downloadable data sets (level 2). Thereby, also a mapping should be made at least to the Blue-Cloud DAB broker common metadata profile as used at level 1 and including use of semantics by controlled vocabularies for several of the key metadata fields.

The following table gives summarised conclusions of the initial analysis per new blue data infrastructure.

Blue Data Infrastructure	**Coupling to Blue-Cloud**	**Conclusions**
EMSO	Direct	EMSO has set up a federated central ERDDAP service, harvesting from local ERDDAP instances for its Observation sites. The deployment of the federation scheme and associated population of the catalogues is progressing and already there is a good basis contents for making the federation in Blue-Cloud. The metadata and data formats of EMSO are well documented and already seem to cover all tags needed for the Blue-Cloud DAB broker metadata profile. Moreover, use is made of controlled vocabularies which will support the DD&AS semantic brokerage approach. Considering level 2, further analysis is needed how to make use of the ERDDAP subsetting ability for querying and retrieving data sets, associated to the collections at level 1. A first approach could be to make no distinction into level 1 and level 2 search and see if downloading provides volume challenges. All data distributed by EMSO ERIC is **Open**, but the specific policy depends on the regional infrastructure. By default, CC-BY is used. Work is in progress at EMSO for deciding on a unified policy for the ERDDAP service output.

Funded by the European Union
Blue-Cloud2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227

Blue Data Infrastructure	Coupling to Blue-Cloud	Conclusions
SIOS	Direct	SIOS is operating a PyCSW based OAI-PMH service, while metadata is based upon the GCMD DIF profile, which is supporting controlled vocabularies. A review is required considering which contents to take onboard. Target could be only directing the SIOS Core Data sets (currently 54) or a subset of the overall 500.000+ records, namely originating from 15 SIOS data providers and with download. Possibilities for level 1 – level 2 distinction and associate mappings, which is preferred considering the large number, will be reviewed, once there is a better understanding of the subcollection and its metadata format. This will also include reviewing the semantic interoperability aspect. All data distributed by SIOS is **Open** and there is a SIOS data policy. No AAI is enforced, but user registration/AAI is required to access higher order services like the basket or transformation services.
EMODnet Physics	Direct	EMODnet Physics is underway with re-organising its ERDDAP services for fitting the Blue-Cloud DD&AS concept by having collection records for specific platforms and associated parameters which are giving access to multiple related data sets records and the actual data sets. The deployment of this re-organisation is ongoing and there is already a test set available with related ERDDAP services. The linking point from level 1 to level 2 is the PLATFORMCODE for each station. The metadata and data formats as used by EMODnet Physics for level 1 and level 2 are well documented and it seems all required Blue-Cloud metadata fields tags are covered, while also use is made of controlled vocabularies to support semantic interoperability. Further analysis is needed if use should be made of the ERDDAP subsetting ability for breaking down long timeseries in smaller data sets. All data distributed by EMODnet Physics are **Open;** by default, CC-BY is used.
ELIXIR-MGnify	Direct	ELIXIR-MGnify service is a relatively new development. It gives results of additional analyses on samples which are described in the ELIXIR-ENA service, which is already federated in the Blue-Cloud DD&AS. There is a MGnify JSON-API for discovery and access. As in the case of ENA, metadata records are the data sets. More analysis is needed for the JSON format to get more insight and options for level 1 – level 2 approach. Possibly, approach of ENA could be adopted. Also, analysis of mapping to the Blue-Cloud DAB common metadata profile and use of controlled vocabularies is still open. The current API supports subsetting functionality, which is implemented by using Elastic for indexing. Most of the data contained in MGnify is open access, under the same licensing model as ENA.

This initial description and evaluation will be followed by a deeper analysis, detailing the way forward for the federation, the possible splitting in level 1 and level 2, the mapping of metadata to the Blue-Cloud common metadata profile for level 1 and additional data for level 2, and the provisions for declaring used controlled vocabularies in order to support the semantic brokerage as planned for the DD&AS. This deeper

analysis will be done by MARIS and CNR-IIA in dialogue with representatives of the new BDIs. Following and as part of the deeper analysis, developments will be undertaken for implementation of the federations. Also, it might be that specific developments will be needed by the BDIs for making their services better fit for Blue-Cloud purpose. These analysis, technical specification, and federation developments for the optimization and expansion of the Blue-Cloud data discovery and access service will be documented in deliverable **D2.3 - Optimised and expanded Blue Cloud Data Discovery and Access Service – Documentation Report,** which is planned for end December 2024.

While in parallel, there will be further analyses and developments ongoing in WP2 with all BDIs – existing and new – for going deeper into subsetting functionalities and establishing and managing data lakes.

The coming deliverables are planned as follows:

- D2.4: BDI sub-setting APIs and Data Lakes – Concept and Specifications Report – end April 2024
- D2.6: Tuning between Blue-Cloud Data Lakes and DTO development, 1st report – end April 2024
- D2.3: Optimised and expanded Blue Cloud Data Discovery and Access Service – Documentation Report – end Dec 2024
- D2.5: Established BDI sub-setting APIs and Data Lakes – Documentation Report – end Feb 2025
- D2.7: Tuning between Blue-Cloud Data Lakes and DTO development, 2nd report – end Feb 2025

Funded by the European Union
Blue-Cloud2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227