

## ***D2.1 Blue Data Infrastructures – Services Description Report***

<b>Work Package</b>	WP2, developing the Blue Cloud discovery and access service and overall Blue Cloud architecture
<b>Lead Partner</b>	MARIS
<b>Lead Author (Org)</b>	MARIS
<b>Contributing Author(s)</b>	Dick M.A. Schaap (MARIS), Peter Thijsse (MARIS), Gilbert Maudire (IFREMER), Cecile Nys (IFREMER), Alain Arnaud (MOI), Renaud Dussurget (MOI), Guy Cochrane (EMBL-EBI), Vishnukumar Balavenkataraman Kadhivelu (EMBL-EBI), Lennert Schepers (VLIZ), Bart VanHoorne (VLIZ), Jean-Olivier Irisson (SU), Benjamin Pfeil (UiB)
<b>Reviewers</b>	Sara Garavelli (TRUST IT), Pasquale Pagano (CNR)
<b>Due Date</b>	31.01.2020, M4
<b>Submission Date</b>	24.02.2020
<b>Version</b>	1.0

### Dissemination Level

- ☒ PU: Public  
☐ PP: Restricted to other programme participants (including the Commission)  
☐ RE: Restricted to a group specified by the consortium (including the Commission)  
☐ CO: Confidential, only for members of the consortium (including the Commission)



## DISCLAIMER

“Blue-Cloud, Piloting Innovative services for Marine Research & the Blue Economy” has received funding from the European Union's Horizon programme call BG-07-2019-2020, topic: [A] 2019 - Blue Cloud services, Grant Agreement n.862409.

This document contains information on Blue-Cloud core activities. Any reference to content in this document should clearly indicate the authors, source, organisation, and publication date.

The document has been produced with the funding of the European Commission. The content of this publication is the sole responsibility of the Blue-Cloud Consortium, and it cannot be considered to reflect the views of the European Commission. The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this document hold any sort of responsibility that might occur as a result of using its content.

## COPYRIGHT NOTICE



This work by Parties of the Blue-Cloud Consortium is licensed under a Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>). “Blue-Cloud, Piloting Innovative services for Marine Research & the Blue Economy” has received funding from the European Union's Horizon programme call BG-07-2019-2020, topic: [A] 2019 - Blue Cloud services, Grant Agreement n.862409.

## VERSIONING AND CONTRIBUTION HISTORY

Version	Date	Authors	Notes
0.1	04.02.2020	MARIS	First version
0.2	18.02.2020	CNR	Internal review
0.3	18.02.2020	TRUST-IT	Internal review
0.4	20.02.2020	MARIS	Final version
0.5	20.02.2020	TRUST-IT	Deliverable shared with the project partners
1.0	24.02.2020	TRUST-IT	Deliverable submission

# Contents

Executive summary.....	4
1 Introduction.....	7
2 Overall concept.....	9
3 Blue Data Infrastructures.....	13
3.1 SeaDataNet .....	13
3.2 EMODnet Bathymetry .....	24
3.3 EMODnet Chemistry .....	30
3.4 EuroArgo - Argo.....	35
3.5 EurOBIS – EMODnet Biology.....	41
3.6 EcoTaxa .....	48
3.7 ELIXIR-ENA.....	52
3.8 EuroBioImaging .....	57
3.9 WEkEO.....	61
3.10 ICOS – Marine.....	65
4 Conclusions and planned follow-up .....	69

## Executive summary

The **Blue Cloud data discovery and access service** will be one of the components of the Blue-Cloud technical framework. It will serve federated discovery and access to blue data infrastructures and interact with the Blue-Cloud Virtual Research Environment (the component federating computing platforms and analytical services). The pilot Blue-Cloud project aims at federating initially in total 10 blue data infrastructures. Each of these existing infrastructures have been described in this deliverable D2.1, in particular with a focus on their current data discovery and access mechanisms.

From the descriptions and discussions, among others at the first TCom meeting in January 2020, it appears that a number of blue data infrastructures do not have to be federated with direct interfacing to the Blue-Cloud data discovery and access service, but indirectly as some are or will be coupled to another of the blue data infrastructures and that way it will be arranged that their contents will also be present in the Blue-Cloud data discovery and access service. In addition, activities are needed in several cases for providing API's which are suited for serving the planned Blue-Cloud data discovery and access service. The following table gives the conclusions of this initial analysis.

Blue Data Infrastructure	Coupling to Blue-Cloud	Conclusions
SeaDataNet	Direct	SeaDataNet operates the Common Data Index (CDI) data discovery and access service. For exchange to Blue-Cloud this already features an INSPIRE compliant API at aggregate metadata level. Still to be specified and developed is a data access API with Marine-ID authentication, capable of processing data requests at aggregate metadata level.
EMODnet Bathymetry	Indirect for data and Direct for products	EMODnet Bathymetry makes use of the SeaDataNet CDI data discovery and access service. Therefore, no separate solutions need to be developed for Blue-Cloud. The highly popular EMODnet Digital Terrain Model (DTM) data product is relevant for Blue-Cloud purposes and can be used through existing OGC web services.
EMODnet Chemistry	Indirect for data and Direct for products	EMODnet Chemistry makes use of the SeaDataNet CDI data discovery and access service. Therefore, no separate solutions need to be developed for Blue-Cloud. The aggregated, harmonized and validated data collections for eutrophication, contamination, acidification and marine litter, as regularly produced by EMODnet Chemistry, are also relevant for Blue-Cloud purposes. Developments are underway for establishing an API and GUI for facilitating sub-setting and retrieval of these data collections. Once operational, this service will provide an additional channel to be added to the Blue-Cloud data discovery and access service.
EuroArgo - Argo	Direct	EuroArgo operates a number of web services for discovery and access to the ArgoFloat data sets. These can be used



Blue Data Infrastructure	Coupling to Blue-Cloud	Conclusions
		for the first release of the Blue-Cloud data discovery and access service. EuroArgo is developing advanced services as part of the ENVRI-FAIR, EOSC-hub, and EA-RISE projects, which should be followed closely as near-future candidate for coupling to the Blue-Cloud.
EuroBIS – EMODnet Biology	Direct	EuroBIS – EMODnet Biology operates a number of web services for discovery and access to the EuroBIS data sets. Of these, the endpoint of the Integrated Publishing Toolkit (IPT) seems to be most suited for connecting to the Blue-Cloud data discovery and access service.
EcoTaxa	Indirect	As part of the Blue-Cloud, EcoTaxa metadata and data will be integrated in EuroBIS – EMODnet Biology by an API which is under development. The coupling to the Blue-Cloud data discovery and access service will then be provided through EuroBIS – EMODnet Biology.
ELIXIR – ENA	Direct	EMBL-EBI operates API's for ENA discovery and ENA data retrieval which seem very suitable endpoints for connecting to the Blue-Cloud data discovery and access service. The ENA system contains many data types / classes and a huge volume of data, which are only partly marine related. Blue-Cloud should focus on data and information relevant for the marine domain and on data types such as samples and their analyses. A priority list needs to be determined as a next step. Moreover, the ENA system offers several algorithms / pipelines for processing data, which might be used in a 'smart' way for the Blue-Cloud. This also need to be analysed.
EuroBioImaging	Direct	As part of EuroBioImaging the BioImage Archive is operated. This consists of 4 separate databases (EMPIAR; Cell-IDR; Tissue-IDR; BioStudies) with different metadata and data models, and different search and access API's. Content is only partly marine related. Blue-Cloud should focus on images and databases relevant for the marine domain. This should be analysed as a next step in order to determine if all databases need to be coupled to the Blue-Cloud.
WEkEO	Direct	WEkEO is under development and will feature the Harmonised Data Access (HDA) API which will allow uniform access to the whole WEkEO catalogue of Sentinel satellite images and Copernicus data products from CMEMS, C3S, CAMS, and CLMS, including subsetting and downloading functionalities. The HDA API will be REST-based. The WEkEO discovery and access services are planned for release in Q1 2020. As soon as launched,

Blue Data Infrastructure	Coupling to Blue-Cloud	Conclusions
		further details for WEkEO need to be gathered and included in the next deliverable D2.2.
ICOS-Marine	Direct	This concerns two relevant portals: ICOS Carbon Portal with data discovery and access and SOCAT portal with data products. Several services are available for both. Some more detail is needed about metadata and data formats, in particular because the ICOS Carbon Portal is upgrading its metadata format and adopting vocabularies as part of the ENVRI-FAIR project. Web services for discovery are existing, but it might be needed to specify and develop API's for data access as part of the Blue-Cloud activities.

In the following months a deeper analysis will be carried out, in order to detail technical specifications of the Blue-Cloud data discovery and access service, and developments required at and by each of the blue data infrastructures. This analysis, technical specification and workplan for the implementation and deployment of the Blue Cloud data discovery and access service will be documented in Deliverable D2.2 at M8. Thereafter, in the months till M17, the actual developments and deployment will take place for establishing an operational Blue-Cloud data discovery and access service by M17.

# 1 Introduction

The technical framework of the pilot Blue-Cloud will feature:

- 1) **the Blue Cloud data discovery and access service** component to serve federated discovery and access to blue data infrastructures
- 2) **the Blue Cloud Virtual Research Environment (VRE)** component to provide a Blue Cloud VRE as a federation of computing platforms and analytical services.

The required data input for the Blue Cloud VRE will be arranged by interaction with the Blue Cloud discovery and access component as well as by users ingesting data sets from other external sources, including their own data sources. The data input can be in-situ data, earth observation data, and model outputs. The analytical services can be various algorithms and generic services, for instance for sub setting, pre-processing, publishing, and viewing data and data products.

The following blue data infrastructures will be pillars under the initial Blue Cloud data discovery and access service:

- SeaDataNet (marine environment) – technically represented in Blue-Cloud by MARIS;
- EMODnet Bathymetry (bathymetry) – technically represented in Blue-Cloud by MARIS;
- EMODnet Chemistry (chemistry) – technically represented in Blue-Cloud by MARIS;
- EuroOBIS – EMODnet Biology (marine biodiversity) – technically represented in Blue-Cloud by VLIZ;
- Euro-Argo and Argo GDAC (ocean physics and marine biogeochemistry)– technically represented in Blue-Cloud by IFREMER;
- ELIXIR-ENA (biogenomics) – technically represented in Blue-Cloud by EBI-EMBL;
- EuroBioImaging (microscopy) – technically represented in Blue-Cloud by EBI-EMBL;
- EcoTaxa (bio images) – technically represented in Blue-Cloud by University of Sorbonne;
- WekEO (CMEMS ocean analysis and forecasting and C3S climate analysis and forecasting) - – technically represented in Blue-Cloud by MOI;
- ICOS-Marine (carbon) – technically represented in Blue-Cloud by University of Bergen.

The Blue Cloud data discovery and access service are analysed and developed in the first 17 months of the project in the following tasks:

- Task 2.1: Developing and deploying the Blue Cloud discovery service (M1 – M17)
- Task 2.2: Developing and deploying the Blue Cloud access service (M4 – M17)

A start was made at the Blue Cloud project kick-off meeting, 2 – 4 October 2019, Pisa – Italy, where the overall concept of the Blue-Cloud technical framework and associated work packages WP2 – WP4 were introduced. Following the kick-off meeting, the technical representatives of the blue data infrastructures as listed above were requested by the WP2 leader (MARIS) to describe their infrastructures, and in particular their data discovery and access mechanisms, in a document following a provided template. This request was met by all representatives and their draft descriptions were analysed by MARIS in search of open issues and/or uncertainties in preparation of the first meeting of the Blue-Cloud TCom. The TCom itself was organized by MARIS as Blue-Cloud Technical Coordinator

in Amsterdam – The Netherlands, 22 – 23 January 2020, to discuss and progress with the technical work packages WP2 – WP4. The first afternoon of the TCom was dedicated to WP2, whereby each of the technical representatives gave a short presentation of their blue data infrastructure, more or less following the earlier provided description. Each presentation was followed by a discussion, whereby more technical questions were posed and answered. Moreover, a part of the afternoon was dedicated to further explaining and discussing the technical approach for the Blue-Cloud data discovery and access service planned, and related activities which will involve all technical representatives of the blue data infrastructures as well as MARIS and CNR-ESSI as developers of the central services and interfaces. In this framework, CNR-ESSI also presented the GEODAB<sup>1</sup> broker which they have developed in earlier projects and which will be instrumental for the planned metadata exchange part. The new insights gained at the TCom meeting together with the earlier descriptive documents have provided the input for this Deliverable D2.1 which focuses on describing the blue data infrastructures and their current services which are of relevance for the Blue-Cloud data discovery and access service.

---

<sup>1</sup> <https://www.geodab.net/>

## 2 Overall concept

The overall concept is that the Blue-Cloud data discovery and access service will harvest metadata from the blue data infrastructures by means of protocols such as CSW or OAI-PMH and using the GEODAB metadata brokerage mechanism (as earlier developed by CNR-IIA). This way, individual outputs will be harmonised to a common metadata model (ISO19115 – 19139), starting at syntactic level. For that purpose, it is planned to make for each of the blue data infrastructures an individual mapping to the common GEODAB metadata model. Thereby also a suitable aggregation should be explored and agreed with each blue data infrastructure to achieve a common Blue-Cloud catalogue with collections of a common level. The resulting Blue-Cloud metadata catalogue will be made available by state-of-art protocols (CSW; OAI-PMH; and SPARQL RDF endpoint), which are directed at machine-to-machine interactions, as well as by means of a GUI for human users. The SPARQL endpoint can be mapped to Schema.org which will ensure a good uptake by major search engines, including the new [Google Data Search engine](#). Schema.org plays a part in the Linked Data approach at the semantic web and is instrumental for publishing structured data on the Internet and to improve the discovery services, by optimising links and codings between different objects optimising findability and access of data for humans, machine-to-machine processes, and leading search engines. Over 10 million sites use Schema.org to markup their web pages and email messages.

For the data access part of the Blue-Cloud data discovery and access service, a data brokerage service will be developed. Therefore, it is planned to analyse for each of the blue data infrastructures their data delivery mechanism. In practice, this might be deployed as a fully open data repository with direct download links (data in different formats and standards). It might also be configured with a shopping mechanism, featuring user login, and possibly making a clear distinction between unrestricted and restricted data.

The planned approach is to adopt or formulate API's together with the technical representatives of the blue data infrastructures. These should be configured by each to be fit for interacting with the central part of the data brokerage service. The API's should deal with the particulars of the local set-ups and also, they should arrange that data requests can be handled and responded at the agreed collection level (see above).

The following images 2.1 – 2.3 illustrate how the Blue-Cloud data discovery and access service will be analysed and developed in steps and also how it will interact with the blue data infrastructures and the Blue-Cloud Virtual Research Environment.

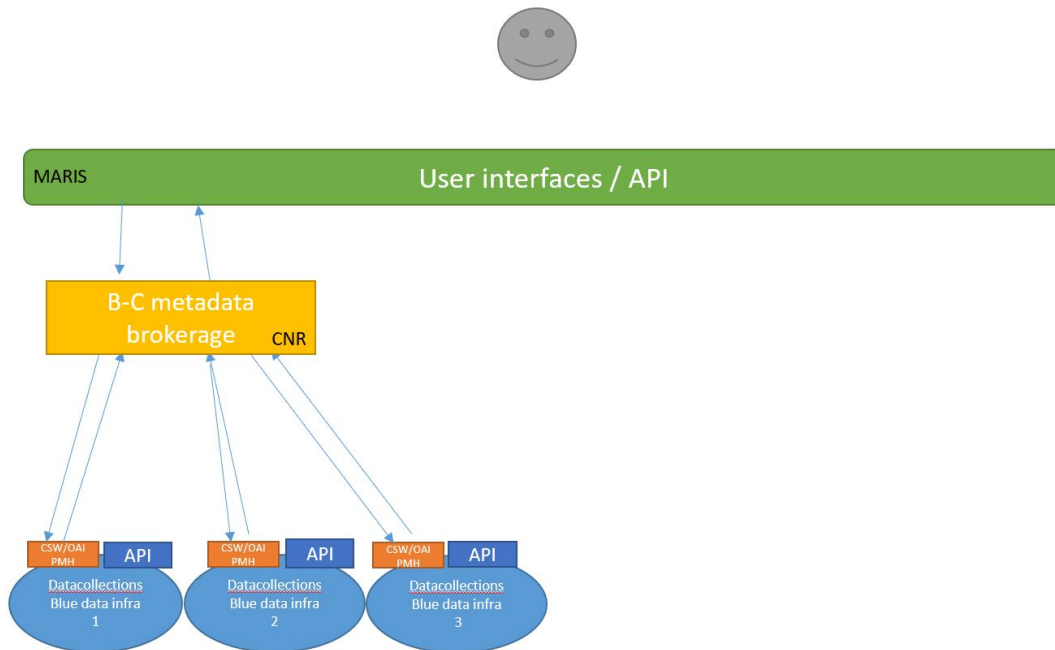


Image 2.1: Indexing and discovery of metadata by users, via a common Blue-Cloud meta catalogue of the metadata brokerage service

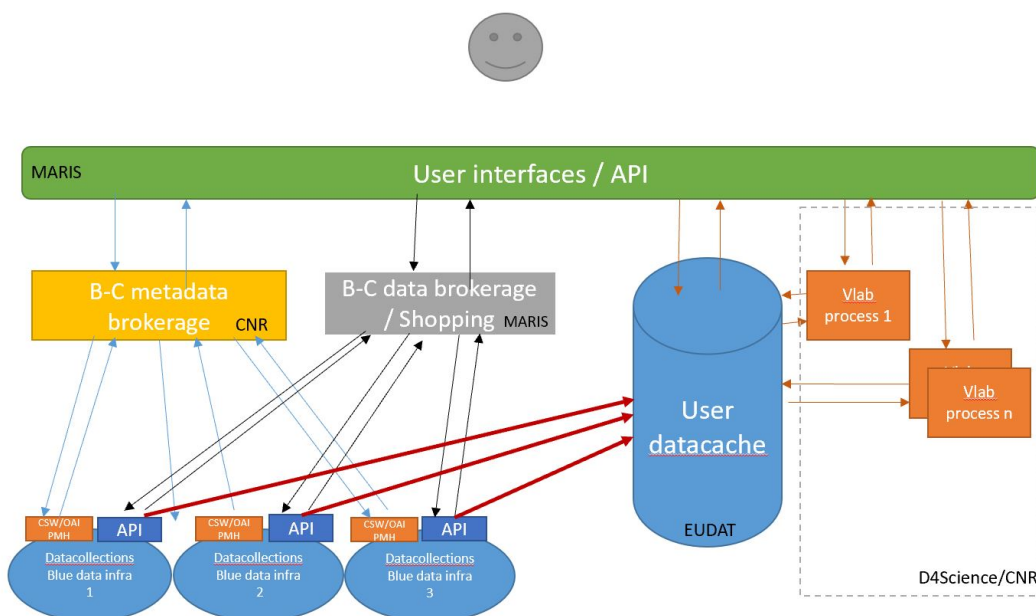
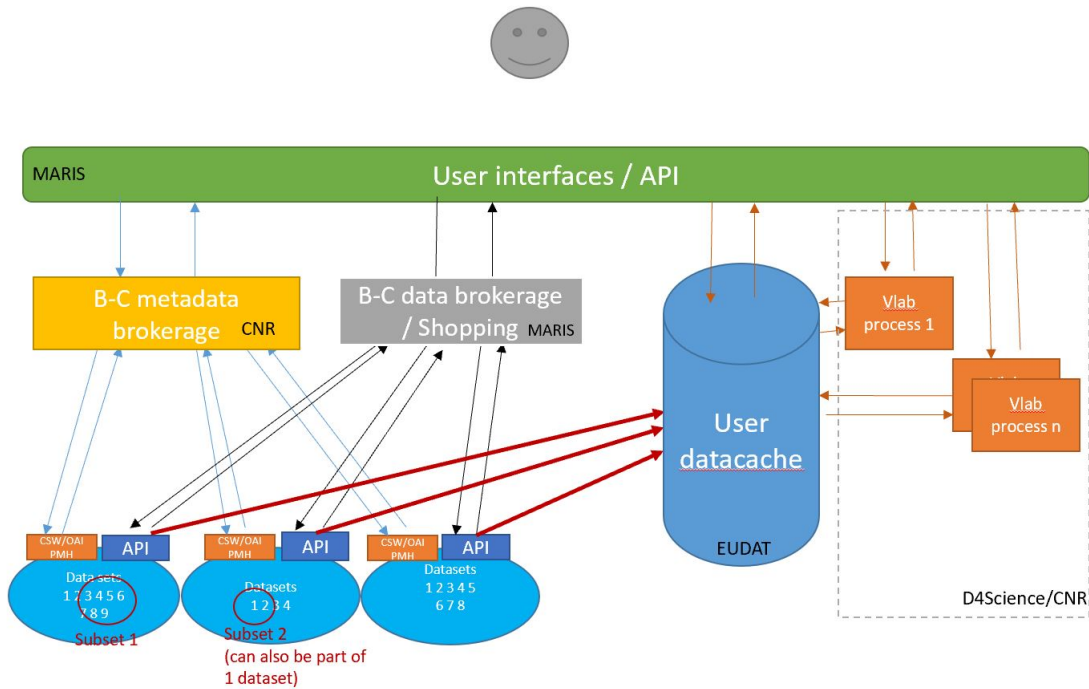


Image 2.2: From metadata to retrieving datasets (as-is) via data brokerage service into user's data cache. Data can be delivered to external users from the data cache. The data cache can also serve as VRE workspace for further processing of the data in Virtual Labs, while the resulting data products can be stored again in the VRE workspace.



*Image 2.3: From metadata to retrieving subsets or processed datasets via data brokerage service and advanced API's at the blue data infrastructures. Resulting data sets again are stored in the data cache and from there available for delivery to external user or part of VRE workspace for further processing in Virtual Labs.*

In this vision, the access links in the Blue-Cloud metadata catalogue will be pointers to the Blue-Cloud data brokerage service, that will interact with the data access API's at the blue data infrastructures. At Blue-Cloud level, the data brokerage will be configured as a shopping mechanism, directly linked to the Blue-Cloud metadata catalogue, whereby individual adaptors will deal with the API's. Moreover, a Blue-Cloud shopping ledger, available for users and providers, will keep track of shopping requests and progress by shopping adaptors for fulfilling the requests. Retrieved datasets will be bundled and made available for downloading from the Blue-Cloud platform. It should be noted that the retrieved data sets will not be harmonized, but in the formats as provided by the individual blue data infrastructures.

This way Blue-Cloud users can discover interesting data collections at the Blue-Cloud central metadata catalogue, followed by requesting access for downloading selected data collections, all through a common Blue-Cloud interface with tracking and tracing.

The Blue-Cloud data discovery and access service will be publicly available for users to download data to their own computers. It will also interact with the Blue Cloud VRE to provide and maintain selected data inputs to the VRE data pool, which can be configured by automatic harvesting following set search profiles.

Initially, the federation as part of the Blue-Cloud data discovery and access service is foreseen as a direct interfacing to the overall collections of data resources as managed by the blue data

infrastructures. But while further analysing and developing technical specifications for the interfaces per blue data infrastructure, there will be situations in which a “smart” federation is more effective and desired. For instance, in case of WekEO, discovery and access are given to the Copernicus collection of Sentinel satellite images. WekEO offers local applications for filtering images, for example on cloud cover, and possibly for geographical subsetting, of which functions use should be made in the Blue-Cloud data discovery and access service. One can also think of a situation that data sets first could be pre-processed at a local blue data infrastructure, using a local algorithm, to reduce their volume and to provide more added-value, before downloading and transferring the retrieved data sets to the user or VRE. In those cases, ‘smart’ federation is needed.

In the present phase of the project, these cases cannot yet be identified. The analysis and development methodology will be based upon a pragmatic principle to go from coarse to fine, starting with the general cases and while doing, identifying cases which should get ‘smart’ solutions.

This stepwise approach and the foreseen interaction of the Blue-Cloud data discovery and access service with the blue data infrastructures and the Blue-Cloud Virtual Research Environment, can be illustrated with the following images.



## 3 Blue Data Infrastructures

In the European landscape of marine and ocean data, great progress has been made in the last two decades with developing standards, services, and establishing dedicated infrastructures for storing, validating, and distributing marine data, and also in a number of cases, for generating and publishing added-value data products. These infrastructures have been and are developed and implemented with support of EU DG RTD (Research and Innovation), EU DG MARE (Maritime Affairs and Fisheries), EU DG GROW (Internal Market, Industry, Entrepreneurship and SMEs), EU DG ENV (Environment), and EU DG CONNECT (Communications Networks, Content and Technology), aiming at developing a European capacity for managing and adding value to marine in-situ and remote sensing data, while federating and interacting with national activities for developing data centers and data management systems, as well as closely interacting with international initiatives.

Generally speaking, the infrastructures are developed and operated by research, governmental, and industry organizations from European states, and closely interacting with international initiatives. They have established links to data originators and their data collection, facilitating to oversee and engage in the process from collection to validation to storage and distribution, while a number of them are also involved in generating data products and knowledge.

The following blue data infrastructures are pillars under the initial Blue Cloud data discovery and access service:

- SeaDataNet (marine environment)
- EMODnet Bathymetry (bathymetry)
- EMODnet Chemistry (chemistry)
- EurOBIS – EMODnet Biology (marine biodiversity)
- Euro-Argo and Argo GDAC (ocean physics and marine biogeochemistry)
- ELIXIR-ENA (biogenomics)
- EuroBioImaging (microscopy)
- EcoTaxa (bio images)
- WekEO (CMEMS ocean analysis and forecasting and C3S climate analysis and forecasting)
- ICOS-Marine (carbon).

These blue data infrastructures are mostly complementary to each other, dealing with other data originators and/or different stages in the processing chains from data acquisition to data products to knowledge. In the following paragraphs, each of them will be described in more details of relevance for the Blue-Cloud data discovery and access service, as foreseen.

### 3.1 SeaDataNet

SeaDataNet (<https://www.seadatanet.org>) is a major pan-European infrastructure for managing, indexing and providing access to marine data sets and data products, acquired by European organisations from research cruises and other observational activities in European coastal marine waters, regional seas and the global ocean. Founding partners are National Oceanographic Data

Centres (NODCs), major marine research institutes, UNESCO-IOC, ICES, and EC-JRC. The SeaDataNet network was initiated in the nineties and over time its network of data centres and infrastructure with standards, tools, and services has expanded, inter alia with support of many EU projects such as Sea-Search, EuroCore, EuMarsin, EuroSeismics, BlackSeaScene, Upgrade-BlackSeaScene, Geo-Seas, MicroB3, and in the last 10 years as part of SeaDataNet, SeaDataNet 2, ODIP 1 & 2, EMODnet projects, and SeaDataCloud. There is close cooperation with various other ocean observing communities such as EuroGOOS, as well as with other major marine data management initiatives and infrastructures, in particular with European Marine Observation and Data network (EMODnet) and Copernicus Marine Environmental Monitoring Service (CMEMS). SeaDataNet develops, governs and promotes common standards, vocabularies, software tools, and services for marine data management, which are freely available from its portal and widely adopted and used. Moreover, the SeaDataNet network of data centres maintains and publishes a series of European directory services which are widely used. These give a wealth of data and information, such as overviews of marine organisations in Europe, and their engagement in marine research projects, managing large datasets, and data acquisition by research vessels and monitoring programmes for the European seas and global oceans.

### **3.1.1 Data discovery and access service component**

A core SeaDataNet service is the **Common Data Index (CDI) data discovery and access service** which provides harmonized discovery and access to a large volume of marine and ocean data sets, both from research and monitoring organisations, which increasingly are major input for developing added-value services and products that serve users from government, research and industry.

#### **3.1.1.1 Name**

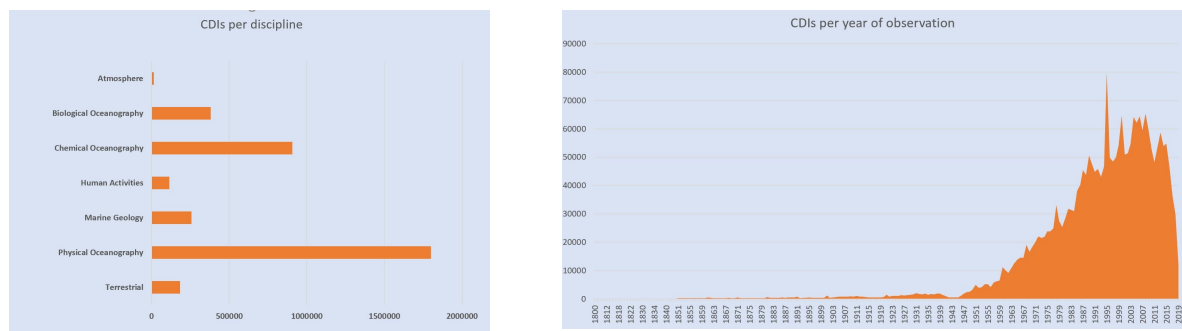
Common Data Index (CDI) data discovery and access service

#### **3.1.1.2 Web address**

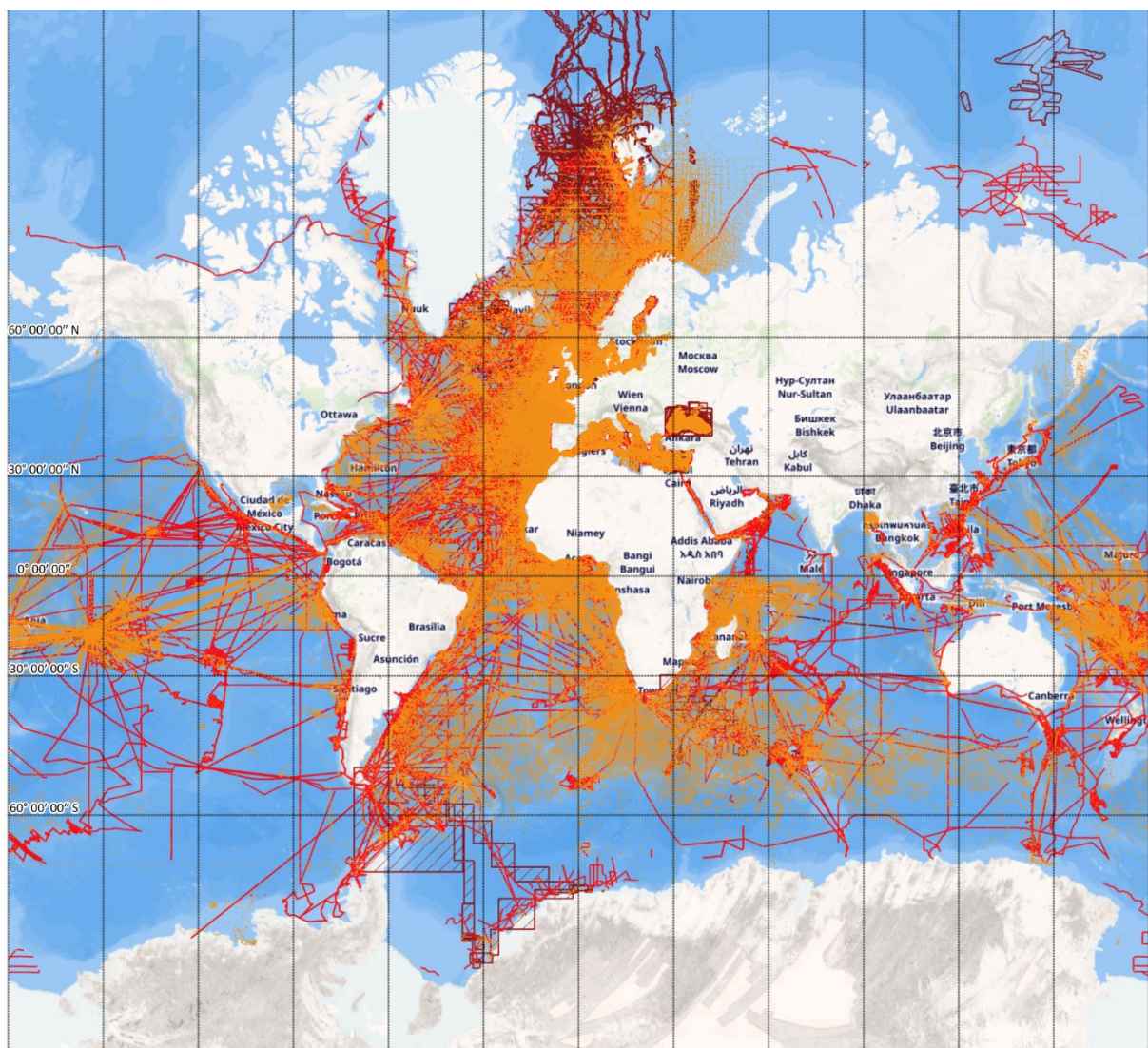
<https://cdi.seadatanet.org/search>

#### **3.1.1.3 Types and number of data sets and/or data products**

The CDI service provides online unified discovery and access to vast resources of data sets, managed by **> 110 connected SeaDataNet data centres from 34 countries** around European seas. Currently it gives access to more than **2.2 Million data sets**, originating from more than **650 organisations** in Europe, covering physical, geological, chemical, biological and geophysical data, and acquired in European waters and global oceans.



*Image 3.1.1: Number of CDI entries per December 2019 per discipline and per year of observation*



*Image 3.1.2: Overview of CDI entries per December 2019*



### 3.1.1.4 Discovery and access mechanisms - how does it function

The online CDI User Interface gives users powerful search options and a highly detailed insight in the availability and geographical spreading of marine data sets, that are managed by the connected data centres. The User Interface includes functions for requesting access, and if granted, for downloading data sets from all connected data centres.

The search function combines free search, facet search and geographic search options, powered by Elastic Search, SQL search, and Geo Server. The data access function comprises a simple and effective data shopping, tracking and download service mechanism. The process from search to getting access to requested data sets is illustrated and explained below.

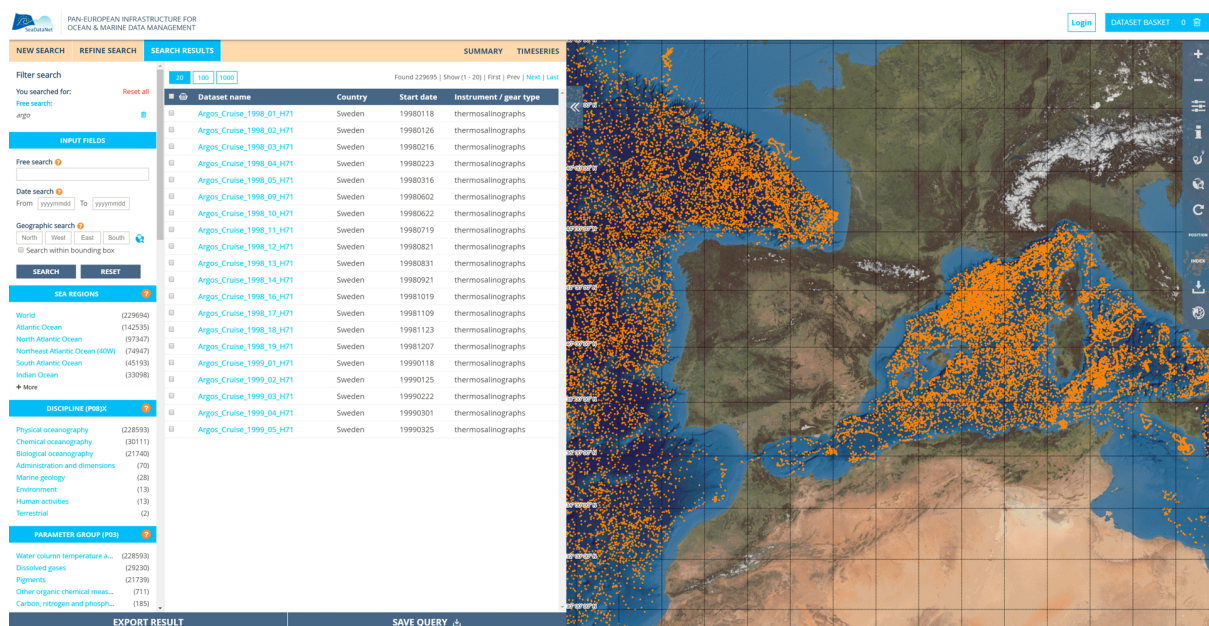


Image 3.1.3: New dynamic user interface of the upgraded CDI data discovery and access service



Image 3.1.4: CDI Shopping Process

The CDI metadata gives information on the what, where, when, how, and who of each data set. It also gives standardised information on the data access restrictions, that apply. The user can include

selected data sets in a shopping basket. All users can freely query and browse in the CDI directory; however, submitting requests for data access via the shopping basket requires that users are registered in the SeaDataNet central user register, thereby agreeing with the overall SeaDataNet User Licence. The registration includes that user receive logon details from the Marine-ID AAI services that SeaDataNet maintains. Data requests can concern unrestricted and/or restricted data sets. Requests for unrestricted data sets are processed immediately after submission and requested data sets are made ready for download automatically from the SeaDataNet central unrestricted data cloud. While requests for restricted data sets are forwarded to the managers of connected data centres for their consideration, most of the cases deliberating with data originators. The processing of all data requests is controlled by the Request Status Manager (RSM) component which is integrated in the CDI User Interface. The RSM registers and processes all transactions, and communicates with the central unrestricted data cloud, users, and data centres. Users receive confirmation e-mails of their data set requests and subsequent processing, and can also check progress and undertake data downloading from the RSM service which is part of their personal dashboard. On their turn, data centres can follow via the RSM service all transactions for their data sets online and can also handle requests for restricted data, which require their mediation. If they agree with restricted data requests, they indicate this in the RSM which triggers that associated restricted data files are uploaded from the local data centre to the user dashboard for downloading. A decision can also be negative. In both cases, users are kept up-to-date by means of daily status emails and by checking their personal dashboard.

The architecture of the latest version of the CDI service is illustrated in the following image. This architecture is the result of an upgrading which has been developed in the ongoing SeaDataCloud project, in a strategic and technical cooperation between the SeaDataNet network and the EUDAT network of academic computing centres. This new set-up has been launched in October 2019 and is fully operational, while further upgrading of the CDI service is undertaken as part of the SeaDataCloud and ENVRI-FAIR projects. The latter is aiming in particular on improving the FAIRness of the service, both by enriching metadata and by optimizing machine-to-machine services. In the new architecture, a separation has been made between the front-end with discovery, shopping and downloading of data sets by users, and the back-end for importing new and updated CDI metadata and related data entries (including versioning) by data centres. The latter is achieved in the new CDI system by introducing a central data cloud, which holds copies of all unrestricted data sets by replication from the connected data centres, and which serves as a central data cache for executing user shopping requests.

Important features of the new CDI system architecture are:

- Data cloud serving as data cache with replicates of unrestricted data sets and with PIDs for improving delivery to users by one data package per shopping basket and without waiting for services of data centres;
- Replication Manager replacing the Download Manager at data centres for ingesting new metadata - data sets and mediating delivery of restricted data sets after approval of data centres;

- Increased quality control of formats and semantics as used by data centres for ingested meta and data and their mutual coherence;
- Data centres make use of an online dashboard which links to the CDI Import Manager service to manage imports of new and updated CDI and data entries. The dashboard also enables data centres to evaluate and process shopping requests for restricted data sets, as well as to oversee and generate reports of all transactions directed at their data centre.
- Import and RSM services work together with the Replication Manager component which is installed and configured at each connected data centre for efficient import and automatic processing. Alternatively, there is a semi-automatic 'interim solution' for data centres that are not allowed to install a local component. Therefore, extra functionality is added to the CDI Import Manager for the CDI central support desk to mediate for 'interim' data centres to maintain the CDI catalogue, both with unrestricted and restricted data sets.

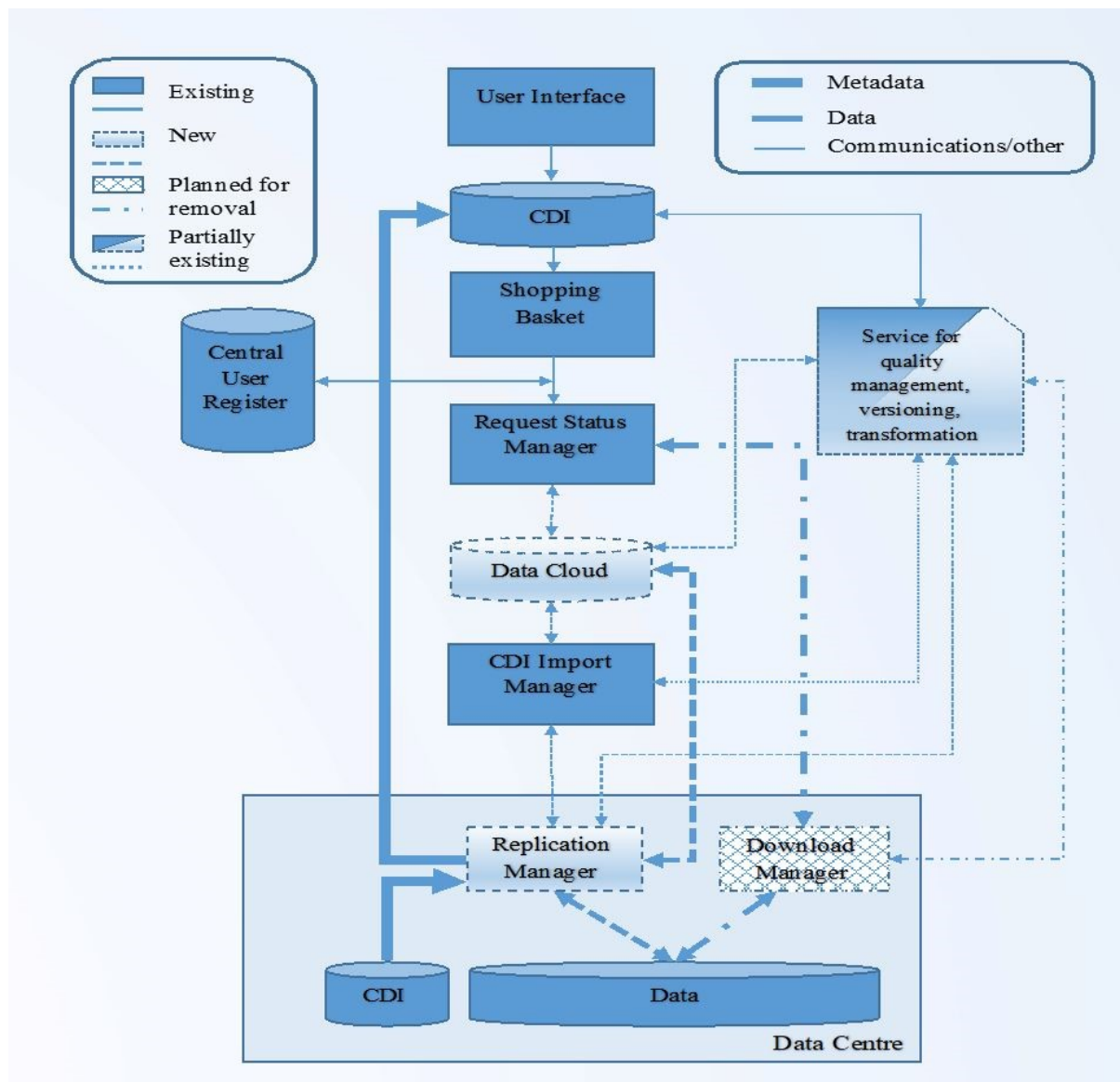


Image 3.1.5: CDI system architecture

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

The **Common Data Index (CDI)** metadata format is based upon the ISO19115 content standard from TC211. The content and XML coding are following the ISO19139 schema and the CDI metadata format is compliant to the INSPIRE Directive Implementing Rules. SeaDataNet maintains and provides to data providers the **MIKADO software tool** to produce CDI XML files from local databases or local metadata files. The CDI metadata format can be considered as a marine profile of the ISO 19115 metadata content standard. The CDI metadata format supports including bibliographic, SeaDataNet Cruise Summary Records (CSR), SeaDataNet Marine Data sets (EDMED), SeaDataNet Research Projects (EDMERP), and quality information. The CDI XML schema implementation is based on the XML schema defined in ISO 19139:2006 TS plus includes additional definitions and Schematron rules. Documentation about the CDI metadata format can be found at: <https://www.seadatanet.org/Standards/Metadata-formats/CDI>

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

Delivery of SeaDataNet data sets to users is done by using, where possible, common SeaDataNet data transport formats, which interact with other SeaDataNet standards such as SeaDataNet controlled vocabularies, SeaDataNet European directories and SeaDataNet Quality Flag Scale as well as with SeaDataNet analysis and presentation tools such as ODV and DIVA. The following SeaDataNet data transport formats have been defined:

- SeaDataNet ODV4 ASCII for profiles, time series and trajectories,
- SeaDataNet NetCDF with CF compliance for profiles, time series and trajectories,
- SeaDataNet MedAtlas as optional extra format,
- NetCDF with CF compliance for gridded data sets

The first 3 formats have been extended with a SeaDataNet semantic header. The **ODV4 format** can be used directly in the popular Ocean Data View (ODV) analysis and presentation software package, which is maintained and regularly extended with new functionalities. The **SeaDataNet NetCDF (CF) format** for profiles, time series and trajectories has been defined by bringing together a community comprising NetCDF and CF experts (such as from NCAR and UNIDATA), and many users of oceanographic point data. This NetCDF format can be used as alternative for the SeaDataNet ODV 4 ASCII format, for profiles, time series and trajectories.

SeaDataNet maintains and provides to data providers the **NEMO software tool** to convert from any type of ASCII format to the SeaDataNet ODV and Medatlas ASCII formats as well as the SeaDataNet NetCDF (CF) format (for timeseries, profiles and trajectories observations) which are then made accessible through the CDI service. Another SeaDataNet **OCTOPUS software tool** is used by data providers as multi-format Checker, Converter and Splitter tool.

Next to these SeaDataNet common data formats, also a number of special SeaDataNet data formats are used and documented for specific data types, such as:

- SeaDataNet ODV ASCII format for **biodiversity data** as developed with EurOBIS;

- SeaDataNet ODV ASCII format for **micro litter data** as developed with EMODnet Chemistry and TG-ML;
- ASCII data format for **beach litter data** as developed with OSPAR, EMODnet Chemistry and TG-ML;
- ASCII data format for **seafloor litter data** as developed with ICES, EMODnet Chemistry and TG-ML;
- SeaDataNet ODV ASCII format for **flowCytoMetry data**;
- NetCDF4 (CF) format for **HF Radar data** as developed with EuroGOOS and EMODnet Physics.

In addition, other common standards can be used for data formats. All applicable SeaDataNet data formats are included in the L24 controlled vocabulary.

Documentation about the SeaDataNet data formats can be found at:  
<https://www.seadatanet.org/Standards/Data-Transport-Formats>

### 3.1.1.7 Use of controlled vocabularies - which, where, how

Use of common vocabularies in all metadatabases and data formats is an important prerequisite towards consistency, interoperability, and FAIRness. Common vocabularies consist of lists of standardised terms that cover a broad spectrum of disciplines of relevance to the oceanographic and wider community. Using standardised sets of terms solves the problem of ambiguities associated with data markup and also enables records to be interpreted by computers. This opens up data sets to a whole world of possibilities for computer aided manipulation, distribution and long-term reuse. Therefore, common vocabularies were set-up and populated by SeaDataNet. The vocabulary services are technically managed and hosted by the British Oceanographic Data Centre (BODC) by means of the NERC Vocabulary Server (NVS2.0). Content governance of the vocabularies is very important and is done by a combined SeaDataNet and MarineXML Vocabulary Content Governance Group (SeaVoX), moderated by BODC, and including many European and international experts.

In addition, the SeaDataNet network maintains and publishes a number of directories:

- European Directory of Marine Organisations (EDMO)
- European Directory of Marine Environmental Data Sets (EDMED)
- European Directory of Marine Research Projects (EDMERP)
- European Directory of Ocean-Observing Systems (EDIOS)
- Cruise Summary Reports (CSR)

These Directories can be linked and referred in metadata and data files, giving additional information, and together with the CDI metadata following a linked data model.

Both the SeaDataNet controlled vocabularies and SeaDataNet Directories are made available as web services for machines and by means of client interfaces for end-users. The client interfaces provide end-users options for searching, browsing and CSV-format export of selected entries. The machine interfaces are provided via a SOAP Application Programming Interface (API) for exchanging



structured information across computer networks as the result of calls. It relies upon XML (eXtensible Markup Language) documents for passing messages. Furthermore, developments are progressing as part of striving for FAIRness for providing each of these vocabularies and Directories also by means of operational SPARQL endpoint for machine interaction.

Documentation about the SeaDataNet controlled vocabularies can be found at:

<https://www.seadatanet.org/Standards/Common-Vocabularies>

Documentation about the SeaDataNet Directories can be found at:

<https://www.seadatanet.org/Metadata>

### 3.1.1.8 Data access policy - if yes, which and how deployed

SeaDataNet has defined and applies an overarching **SeaDataNet Data policy**, that aims to strike a balance between the rights of investigators and the need for widespread access through the free and unrestricted sharing and exchange of SeaDataNet data, meta-data and data products. The final goal of this policy is to serve the scientific community, public organisations, and environmental agencies, and to facilitate the production of advice and status reports by stating the conditions for data submission, access and use. This policy applies to data managed by SeaDataNet partners for providing access to data managed in the SeaDataNet CDI service. Part of the SeaDataNet Data Policy is the SeaDataNet User Licence, to which every user agrees as part of the process to register as a SeaDataNet User and to get Marine-ID logon details which enable to submit CDI data access requests, to download CDI related data, and to follow progress of data requests in the CDI RSM service.

Documentation about the SeaDataNet Data Policy and SeaDataNet User Licence can be found at:

<https://www.seadatanet.org/Data-Access/Data-policy>

### 3.1.1.9 Any web services and API's - URLs, function, how to operate

SeaDataNet maintains SOAP and SPARQL web services for the controlled vocabularies and Directories. For the CDI service, SeaDataNet operates a web service which provides aggregated collections of CDIs at: <https://cdi.seadatanet.org/report/aggregation>

The current CDI aggregation works by 3 factors, namely:

- Active combinations of organization codes (=EDMO codes) for CDI-author\_Data-Custodian\_Data-Distributor;
- Active area-types (=L02 codes) for Point/Curve/Surface;
- Active Parameter Disciplines (=P08 codes)

This way the more than 2.3 million CDI records result in circa 1000+ aggregated CDI records, which make use of the same CDI INSPIRE compliant metadata schema.

These aggregated CDI XML records are already harvested dynamically by the GEODAB broker service, operated by CNR-ESSI, which results in CDI aggregate end-points following the OGC CSW, OGC OpenSearch, and OAI-PMH protocols. The URLs for the CDI service endpoints are as follows:

SeaDataNet OGC CSW endpoint:

<http://gs-service-production.geodab.eu/gs-service/services/essi/view/seadatanet/csw>

with SeaDataNet CSW GetCapabilities:

<http://gs-service-production.geodab.eu/gs-service/services/essi/view/seadatanet/csw?service=CSW&request=GetCapabilities&version=2.0.2>

SeaDataNet OAI-PMH endpoint:

<http://gs-service-production.geodab.eu/gs-service/services/essi/view/seadatanet/oaipmh>

with SeaDataNet OAI-PMH Identify request:

<http://gs-service-production.geodab.eu/gs-service/services/essi/view/seadatanet/oaipmh?verb=Identify>

OGC OpenSearch description endpoint:

<http://gs-service-production.geodab.eu/gs-service/services/essi/view/seadatanet/opensearch/description>

With OpenSearch based demo portal:

<http://gs-service-production.geodab.eu/gs-service/search?view=seadatanet>

These web services are being harvested by the GEOSS portal and the Ocean Data Portal of IOC-IODE.

Furthermore, activities are underway for setting up a SPARQL endpoint for the CDI service.

#### **3.1.1.10 Any suggestions for aggregating data sets as collections in order to scale down number of too many entries and serve users with distinctive metadata entries as part of the planned Blue-Cloud data discovery and access service**

For the Blue-Cloud data discovery and access service, use can be made of the existing CDI aggregate endpoint. However, the existing aggregation can result in CDI collections which are in several cases very large. For the shopping mechanism SeaDataNet maintains a maximum number of 10.000 data sets per shopping basket, while multiple baskets are allowed. Therefore, the data access API (still to be developed) should take this into account, possibly by ‘chopping’ large requests from the CDI aggregates into multiple smaller basket requests.

#### **3.1.1.11 Hosting environment**

The main components of the CDI service are hosted at MARIS, IFREMER, BODC, HCMR, and EUDAT. MARIS hosts:

- CDI User Interfaces (for humans and machines)
- CDI Metadatabase
- CDI Shopping front-end and Request Status Manager
- CDI Import Manager, including quality control on CDI syntax and semantics, coherence with data files, and versioning

These components steer the front-end process of discovery, shopping, and delivery, and the back-end process of import, validation, and publishing.

IFREMER hosts:

- Marine-ID AAI service for registration and authentication
- SeaDataNet User Register

IFREMER also maintains the CDI software tools: MIKADO, NEMO, OCTOPUS, and Replication Manager

BODC hosts:

- SeaDataNet Controlled Vocabularies
- SeaVox governance board for vocabularies

HCMR hosts:

- Monitoring portal which facilitates alerts and reports on availability and performance of the CDI service and each of its components

EUDAT hosts:

- Central Data Cloud for unrestricted data files
- Component for central quality checks on syntax and semantics of data files
- Component for giving PIDs to imported data files
- Component for handling import of data files
- Component for handling retrieval and delivery of data files
- Component for monitoring the availability and functioning of the CDI service components

The processes at EUDAT are steered by the Shopping and Import modules at MARIS.

Data Providers host:

- Replication Manager to interact with the CDI Import Manager for updates and new entries of CDI metadata and associated data files, and with the Request Status Manager in case of agreed requests for restricted data files
- Metadata and data sets as received from data originators, which have been elaborated and validated for SeaDataNet exchange.

The following image gives an overview of the SeaDataNet CDI hosting configuration:

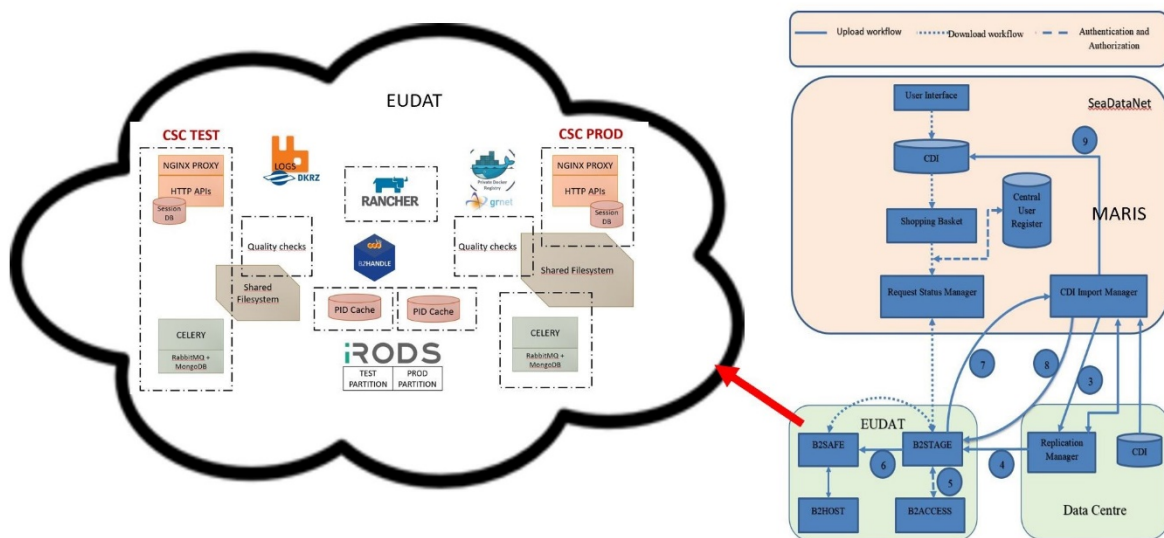


Image 3.1.6: Hosting environment of the CDI service components with EUDAT part highlighted

#### **3.1.1.12 Organisational aspects (main operator(s); data providers)**

Main operators are: MARIS, IFREMER, BODC, HCMR, and EUDAT members (CSC, CINECA, GRNET, DKRZ, and STFC). Data Providers: currently >110 data centres consisting of NODCs, marine research institutes, hydrographic surveys, geological institutes, environmental monitoring agencies, international organisations, and some industry.

#### **3.1.1.13 Contact details**

CDI main operator contact: MARIS, The Netherlands, Dick M.A. Schaap – dick@maris.nl

#### **3.1.1.14 Conclusion SeaDataNet**

SeaDataNet operates the CDI data discovery and access service. For exchange to Blue-Cloud this already features an INSPIRE compliant API at aggregate metadata level. Still to be specified and developed is a data access API with Marine-ID authentication, capable of processing data requests at aggregate metadata level.

### **3.2 EMODnet Bathymetry**

The European Marine Observation and Data network (EMODnet) (<https://www.emodnet.eu>) was initiated in 2008 and it is a long-term, marine data initiative funded by the European Maritime and Fisheries Fund (managed by EU DG MARE), which, together with the Copernicus space programme and the Data Collection Framework for fisheries, implements the EU's Marine Knowledge 2020 strategy. EMODnet connects a network of over 150 organisations supported by the EU's Integrated Maritime Policy who work together to observe the sea, process the data according to international standards and make that information freely available as interoperable data layers and data products. This 'collect once and use many times' philosophy benefits all marine data users, including policy makers, scientists, private industry and the public. It has been estimated that this kind of integrated marine data policy will save off-shore operators at least one billion Euro per year, as well as opening up new opportunities for innovation and growth. The aim of EMODnet is to increase productivity in all tasks involving marine data, to promote innovation and to reduce uncertainty about the behavior of the sea. This will lessen the risks associated with private and public investments in the blue economy, and facilitate more effective protection of the marine environment.

EMODnet provides easy and free access to marine data, metadata and data products and services spanning seven broad disciplinary themes: bathymetry, geology, physics, chemistry, biology, seabed habitats and human activities. Each theme is dealt with by a partnership of organisations that possess the expertise necessary to standardise the presentation of data and create data products. Moreover, for each of the portals use is made of existing data management infrastructures, which are dealing with bringing data originators and data together, and which are providing relevant base data for developing EMODnet products and derived services. The synergy with EMODnet also provides a boost to the existing data management infrastructures as more data providers are stimulated to participate and share their data for EMODnet products. EMODnet turns marine data into maps, digital terrain

models, time series & statistics, dynamic plots, map viewers and other applications ready to support researchers, industries and policy makers to tackle grand societal challenges.

EMODnet currently offers the following thematic portals:

- EMODnet Bathymetry
- EMODnet Chemistry
- EMODnet Physics
- EMODnet Biology
- EMODnet Seabed Habitat mapping
- EMODnet Geology
- EMODnet Human Activities

And two common portals:

- EMODnet Central portal
- EMODnet Ingestion portal

### **3.2.1 EMODnet Bathymetry details**

The EMODnet Bathymetry portal is operated and further developed by a European partnership. This comprises members of the SeaDataNet consortium together with organisations from marine science, the hydrographic survey community, and industry. The partners combine expertises and experiences of collecting, processing, and managing of bathymetric data together with expertises in distributed data infrastructure development and operation and providing OGC services (WMS, WFS, and WCS) for viewing and distribution.

The main aims of EMODnet Bathymetry are:

- To bring together available bathymetric surveys and derived high-resolution composite DTMs (Digital Terrain Models)
- To produce and maintain the best Digital Terrain Model for the European seas based upon the gathered bathymetry data and with a grid resolution of 1/16 arc minute \* 1/16 arc minute (circa 115 meter \* 115 meter)
- To publish and disseminate the EMODnet DTM widely with metadata, acknowledging used data and their data providers, OGC viewing services, and download services

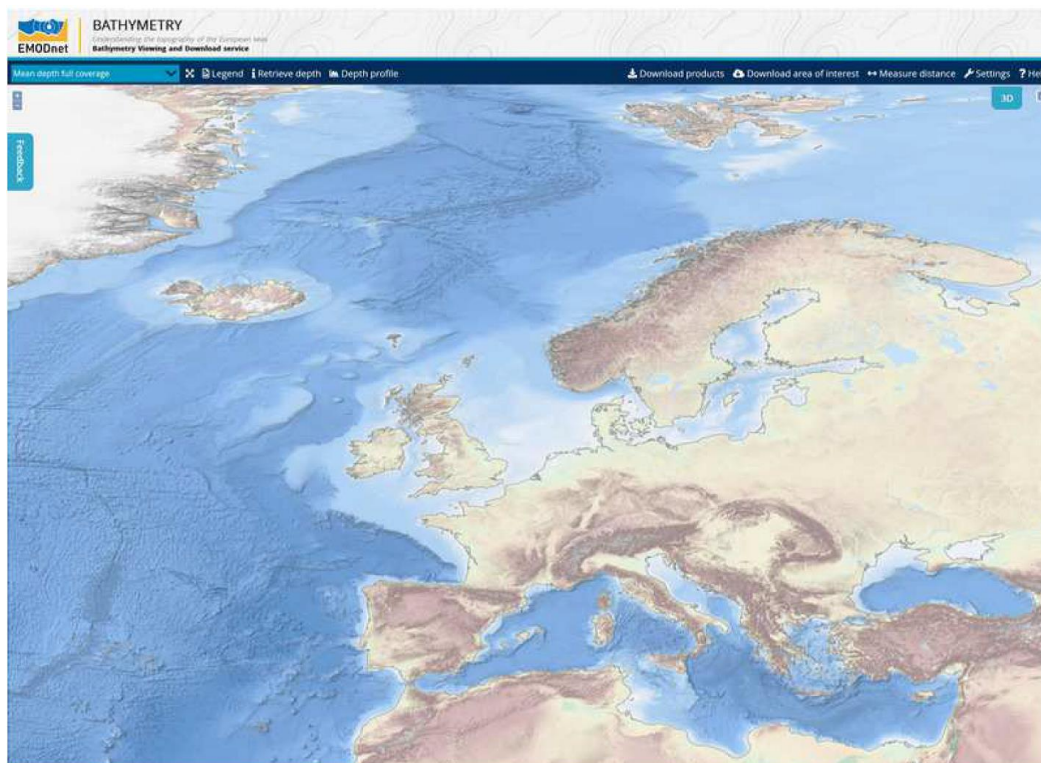
A methodology is followed that data providers populate their data sets with metadata in a general catalogue and thereafter, pre-process and grid their data sets with a standard software tool (GLOBE) to a preset grid resolution and gridding for handover as input for Regional Coordinators (RCs). These RCs use the received input to generate a Regional DTM for each of the defined European sea regions, using the GLOBE software, and at a target grid resolution. Thereafter, Regional DTMs are compiled together into an integrated EMODnet DTM, thereby performing a lot of efforts on quality control, identification and corrections of anomalies, and ensuring coherent references to surveys and cDTMs

used and GEBCO. This is followed by making the DTM ready for publishing as viewing and download services.

The current EMODnet DTM version has been released officially on 24th September 2018. It contains approx. 12.3 billion grid nodes, organized in 113892 columns and 108132 rows (seabed and terrestrial coverage included) while e.g. GEBCO has 933 million grid nodes for worldwide coverage. From all the data sources gathered, a total of 9369 unique survey references and 87 cDTM references are used in the overall DTM. Also Satellite Derived Bathymetry (SDB) data products have been included, in particular for coastal zones. While, gaps in coverage have been completed by using GEBCO. The EMODnet DTM is available free of charge for viewing and downloading, and sharing by OGC web services from the EMODnet Bathymetry portal.

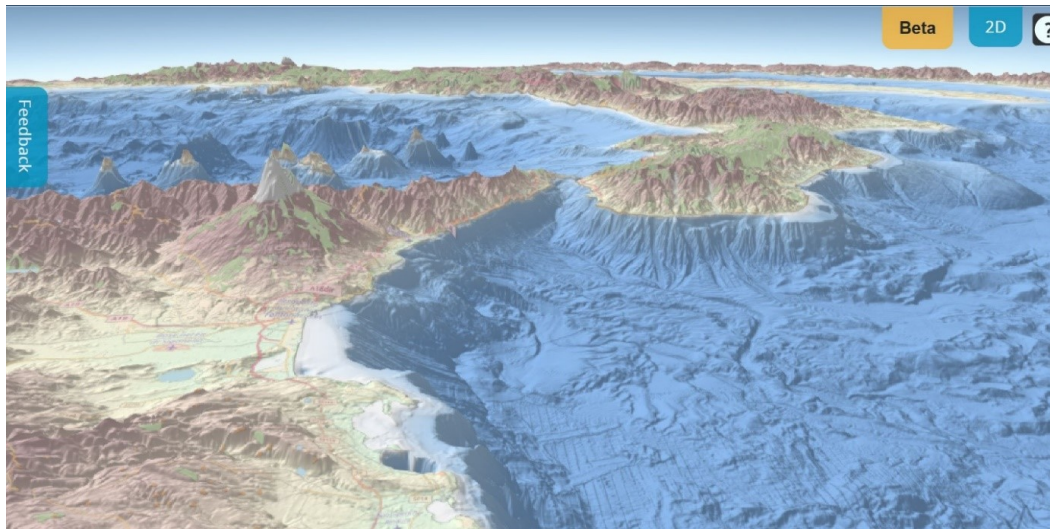
Other features of the latest release are:

- A powerful 3D bathymetry visualisation functionality in the viewer without plugins;
- All European seas including part of the Arctic Ocean and Barents Sea;
- Inclusion of Satellite Derived Bathymetry data products, in particular for coastal stretches of Spain and Greece;
- Improved source reference layer with quality indication;
- Downloading of DTM tiles is integrated into a shopping mechanism which facilitates registration of users and their reasons for use.



*Image 3.2.1: New EMODnet DTM with higher resolution and including arctic waters*





*Image 3.2.2: View from the South along Sicily and the Southern part of mainland Italy*

The GIS layers in the Bathymetry Viewing and Download service can be shared as OGC WMS and WCS services with other EMODnet portals and beyond. Also, WMS layers from other EMODnet portals and external services can be added to the Bathymetry Viewer and Download service. The URLs for the OGC services can be found at the portal.

The portal and the EMODnet DTM are very successful with many users from research, government and industry. EMODnet Bathymetry is cited on multiple occasions as a successful regional project involving both hydrographic offices and research institutes with complementing approaches in terms of data coverage and methodologies (acquisition, processing and validation). The web portal and its services are well visited with > 10.000 unique visitors per month. The OGC web services (machine-to-machine) are very popular with more than 250.000 visitors per year. The number of downloaded DTM tiles amounts to circa 10.000 per 3 months.

The current EMODnet Bathymetry phase runs till end 2020 with option for a seamless 2 years continuation and it is undertaken by a consortium of 41 partners, led by SHOM (coordinator) and MARIS (technical coordinator). Its major aim is to refine the EMODnet DTM further with additional high-quality data sets and improved methodologies.

### **3.2.2 Data discovery and access service component**

EMODnet Bathymetry makes full use of the SeaDataNet infrastructure for managing the gathering of bathymetry data sets. References to the used data and their data holders can be found in the source references layer. Gathered survey datasets are described and included in the SeaDataNet Common Data Index (CDI) Data Discovery and Access service, metadata about composite DTMs are included in the SeaDataNet Sextant Catalogue service for data products.

#### **3.2.2.1 Name**

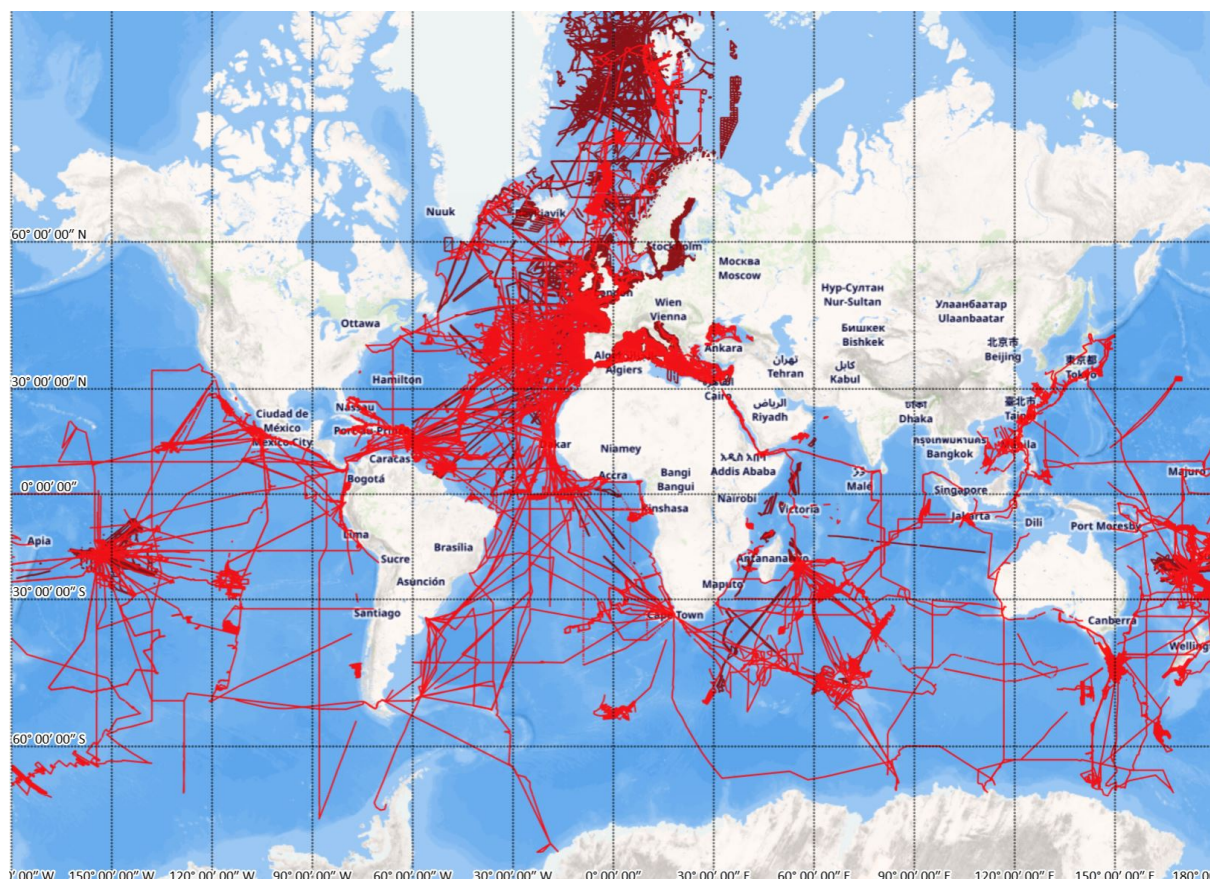
Bathymetry Common Data Index (CDI) data discovery and access service

### 3.2.2.2 Web address

<https://www.emodnet-bathymetry.eu/search>

### 3.2.2.3 Types and number of data sets and/or data products

The EMODnet Bathymetry subset of the CDI service provides online unified discovery and access to bathymetry survey data sets from single beam, multibeam and other surveys. Currently it comprises circa 27.000 entries, brought together by **42 data centres from 250 originators**.



*Image 3.2.3: Overview of EMODnet Bathymetry CDI entries per December 2019*

The highly popular EMODnet Digital Terrain Model (DTM) can be used for Blue-Cloud purposes by its OGC web services:

WMS: <https://ows.emodnet-bathymetry.eu/wms>

WFS: <https://ows.emodnet-bathymetry.eu/wfs>

WMTS: <https://tiles.emodnet-bathymetry.eu>

WCS: <https://ows.emodnet-bathymetry.eu/wcs>

### 3.2.2.4 Discovery and access mechanisms - how does it function

See description for SeaDataNet CDI data discovery and access service



### **3.2.2.5 Metadata format(s) - short overview and references to detailed documentation**

See description for SeaDataNet CDI data discovery and access service

### **3.2.2.6 Data format(s) - short overview and references to detailed documentation**

See description for SeaDataNet CDI data discovery and access service. Use is made of SeaDataNet ODV4, BAG, ESRI XYZ, and NetCDF (CF) formats.

### **3.2.2.7 Use of controlled vocabularies - which, where, how**

See description for SeaDataNet CDI data discovery and access service.

### **3.2.2.8 Data access policy - if yes, which and how deployed**

See description for SeaDataNet CDI data discovery and access service.

### **3.2.2.9 Any web services and API's - URLs, function, how to operate**

See description for SeaDataNet CDI data discovery and access service.

### **3.2.2.10 Any suggestions for aggregating data sets as collections in order to scale down number of too many entries and serve users with distinctive metadata entries as part of the planned Blue-Cloud data discovery and access service**

See description for SeaDataNet CDI data discovery and access service.

### **3.2.2.11 Hosting environment**

See description for SeaDataNet CDI data discovery and access service

### **3.2.2.12 Organisational aspects (main operator(s); data providers)**

See description for SeaDataNet CDI data discovery and access service. Data Providers: currently >40 data centres consisting of NODCs, marine research institutes, hydrographic surveys, and some industry

### **3.2.2.13 Contact details**

CDI main operator contact: MARIS, The Netherlands, Dick M.A. Schaap – dick@maris.nl

### **3.2.2.14 Conclusion EMODnet Bathymetry**

EMODnet Bathymetry makes use of the SeaDataNet CDI data discovery and access service. Therefore, no separate solutions need to be developed for Blue-Cloud.

The highly popular EMODnet Digital Terrain Model (DTM) is also relevant for Blue-Cloud purposes and can be used through existing OGC web services.

### 3.3 EMODnet Chemistry

EMODnet Chemistry is also a thematic EMODnet portal. For more background on EMODnet, see EMODnet Bathymetry.

#### 3.3.1 EMODnet Chemistry details

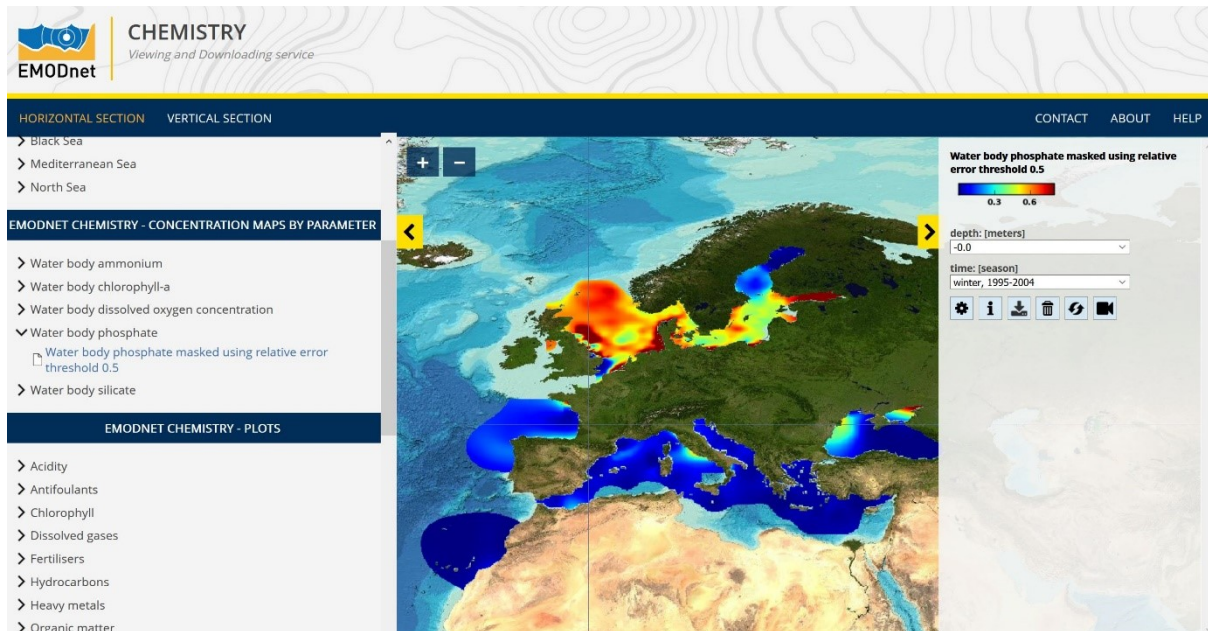
The EMODnet Chemistry portal is operated and further developed by a European partnership. This comprises members of the SeaDataNet consortium together with organisations from marine science, environmental monitoring agencies, regional sea conventions, ICES, EEA, chemical experts, and others. The partners combine expertises and experiences of collecting, processing, and managing of chemistry data together with expertises in distributed data infrastructure development and operation and providing OGC services (WMS, WFS, and WCS) for viewing and distribution.

The main aims of EMODnet Chemistry are:

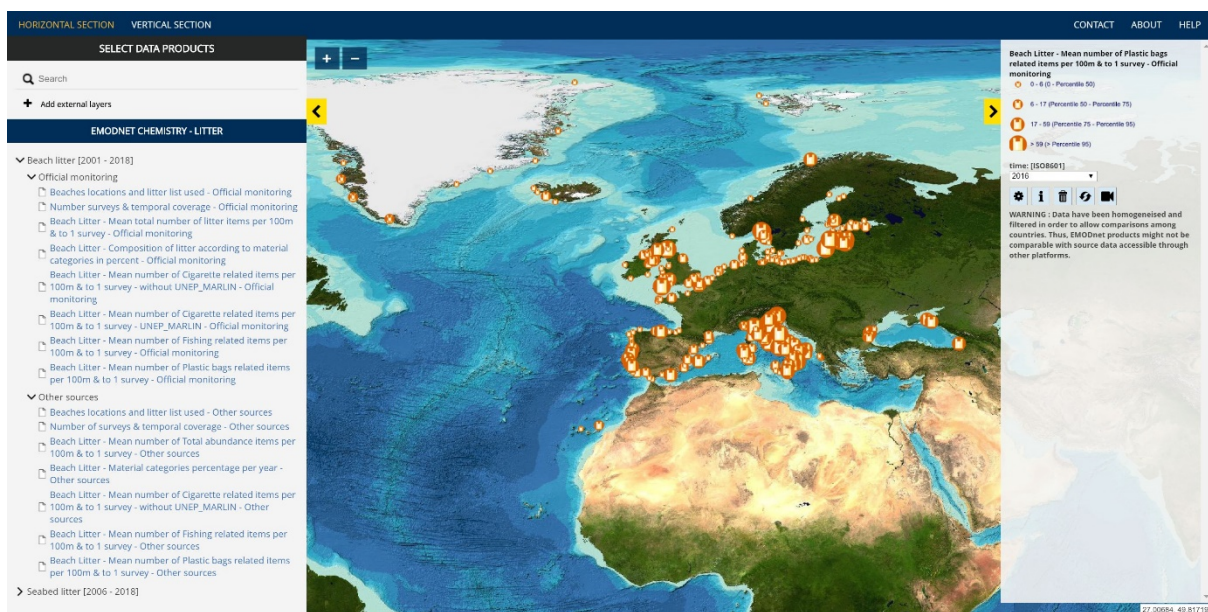
- To bring together available chemistry observation data for eutrophication, contaminants and marine litter
- To produce and maintain validated aggregated and harmonised data collections and interpolated map products for eutrophication, contaminants and marine litter, fit for purpose for support of implementation of the Marine Strategy Framework Directive (MSFD)
- To publish and disseminate the EMODnet Chemistry data products widely with metadata, acknowledging used data and their data providers, OGC viewing services, and download services.

Data gathering in EMODnet Chemistry is done in direct communication with data originators to ensure the best sets of measured data and related metadata, and to prevent duplicates. The gathered data are aggregated and validated by MSFD region. Thereby a major challenge is to manage the heterogeneity, complexity, quality and large volume of the gathered datasets and to process these into harmonized data collections. This is solved by using consolidated SeaDataNet standards for vocabularies, QA-QC, and software tools. This activity results in harmonized validated data collections for each MSFD region, concerning eutrophication (MSFD indicator 5) and contaminants (MSFD indicators 8 and 9). These data collections are input for generating further data products, consisting of a series of spatially interpolated maps of eutrophication parameters in time and depth per sea region, and station time series of contaminants parameters. The resulting products are published for users for browsing, interacting and visualizing by means of dedicated viewing services, while metadata of the data products can be retrieved from a products catalogue service. For marine litter (MSFD indicator 10), the focus is on beach litter, seafloor litter and micro plastics. Data for beach litter and sea floor litter are gathered, managed and published by means of two central databases, which are developed and populated by EMODnet Chemistry in cooperation with the MSFD Technical Group on Marine Litter (TG-ML), EU JRC, RSC's, ICES, and several relevant

EU projects, regional and local initiatives. For micro plastics the SeaDataNet CDI service approach has been adapted and dedicated guidelines have been formulated and published in concertation with the TG-ML.



*Image 3.3.1: OceanBrowser service displaying interpolated seabasin map for water body phosphate*



*Image 3.3.2: European beach litter map of mean number of plastic bags related items per 100 m*

The products are described with metadata in the Chemistry products catalogue, have DOIs and landing pages for citation, and can be viewed in the OceanBrowser service, while there are also web services for sharing with other portals. These GIS layers in the Chemistry Viewing service can be

shared as OGC WMS service with other EMODnet portals and beyond. Also, WMS layers from other EMODnet portals and external services can be added to the Chemistry Viewer service. The URLs for the OGC services can be found at the portal.

EMODnet Chemistry undertakes close cooperation and tuning with the European Environment Agency (EEA) and the 4 Regional Sea Conventions (OSPAR, HELCOM, Bucharest Convention, and Barcelona Convention), JRC and ICES for making the data products fit for use in the MSFD process, while recently a MoU has been established with Copernicus Marine Environmental Monitoring Service (CMEMS) for exchanging validated data products for eutrophication to CMEMS for further developing their ecosystem modelling and products.

The current EMODnet Chemistry phase runs till spring 2021 with option for a seamless 2 years continuation and it is undertaken by a consortium of 48 partners, led by OGS (coordinator) and MARIS (technical coordinator). Its major aim is to refine the EMODnet Chemistry products further with additional high-quality data sets and improved methodologies.

### **3.3.2 Data discovery and access service component**

EMODnet Chemistry makes full use of the SeaDataNet infrastructure for managing the gathering of chemistry data sets. References to the used data and their data holders can be found in the data products. Gathered chemistry datasets are described and included in the SeaDataNet Common Data Index (CDI) Data Discovery and Access service, while metadata about data products are included in the SeaDataNet Sextant Catalogue service for data products.

#### **3.3.2.1 Name**

Chemistry Common Data Index (CDI) data discovery and access service

#### **3.3.2.2 Web address**

<https://emodnet-chemistry.maris.nl/search>

#### **3.3.2.3 Types and number of data sets and/or data products**

The EMODnet Chemistry subset of the CDI service provides online unified discovery and access to chemistry data sets for eutrophications (nutrients, oxygen, chlorophyll), contaminants, marine litter (beach, seafloor and micro litter). Currently it comprises circa 1 million entries, brought together by **65 data centres from 419 originators**.



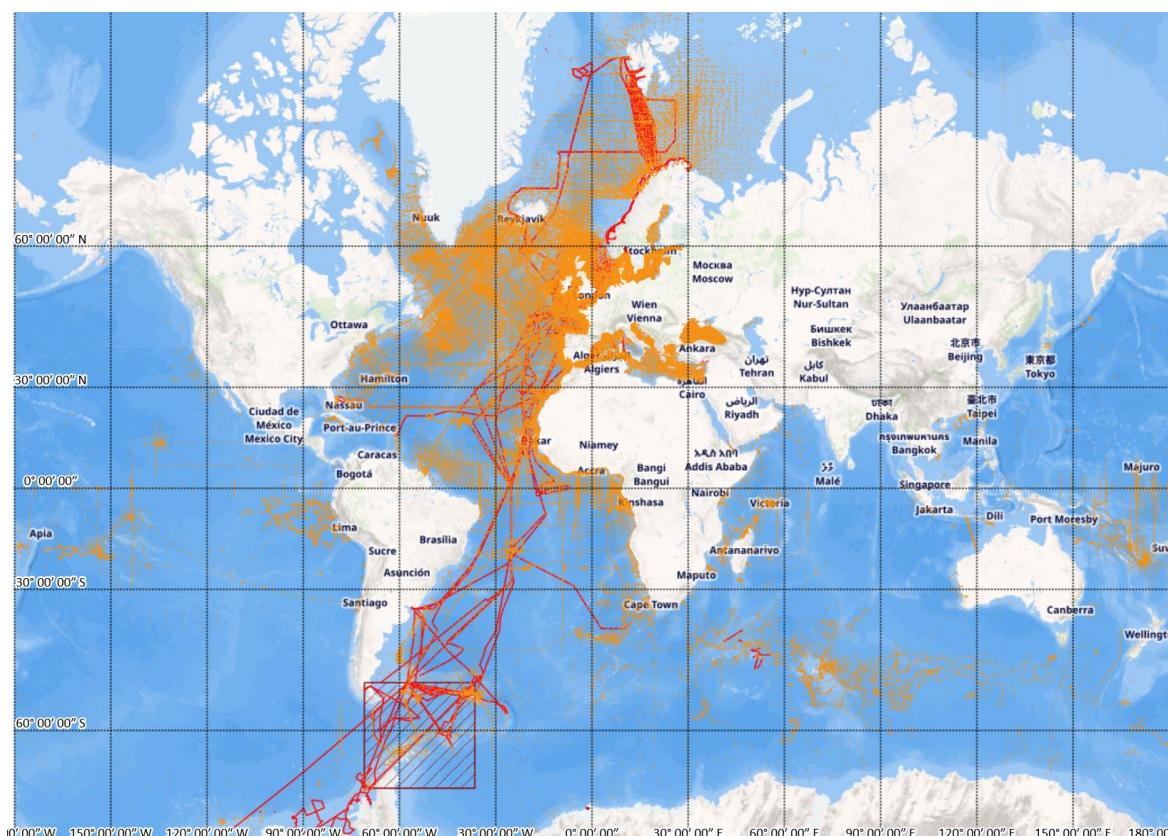


Image 3.3.3: Overview of EMODnet Chemistry CDI entries per December 2019

Group of Variables	Baltic Sea	Iberian peninsula - Macaronesia - Bay of Biscay	Greater North Sea - Celtic Sea - Faroes	Arctic Ocean - Norwegian Sea - Greenland Sea - Barents Sea - Icelandic Waters	Mediterranean Sea	Black Sea - Sea of Marmara - Sea of Azov
Acidity <sup>i</sup>	■	■	■	■	■	■
Antifoulants <sup>i</sup>	■	■	■		■	■
Chlorophyll <sup>i</sup>	■	■	■	■	■	■
Dissolved gasses <sup>i</sup>	■	■	■	■	■	■
Fertilisers <sup>i</sup>	■	■	■	■	■	■
Hydrocarbons <sup>i</sup>	■	■	■	■	■	■
Heavy metals <sup>i</sup>	■	■	■	■	■	■
Marine litter <sup>i</sup>	■	■	■	■	■	■
Organic matter <sup>i</sup>	■	■	■	■	■	■
Polychlorinated biphenyls <sup>i</sup>	■	■	■	■	■	■
Pesticides and biocides <sup>i</sup>	■	■	■	■	■	■
Radionuclides <sup>i</sup>	■		■	■	■	■
Silicates <sup>i</sup>	■	■	■	■	■	■

■ 1-50	■ 251-1000	■ 2501-5000	■ 10001-25000
■ 51-250	■ 1001-2500	■ 5001-10000	■ >25000

Legend - number of measurement data sets for each variable per marine region

Image 3.3.4: Chemistry CDI matrix user interface by chemical substances and marine regions

Important data products of EMODnet Chemistry are EMODnet Chemistry aggregated, harmonized and validated data collections. These are made by regularly processing the CDI data entries for European seas and for specific chemical substances, such as eutrophication, contaminants, acidification, and marine litter. The processing includes validation, harmonization, and aggregation of the data entries from more than 60 data centres. These collections are added-value data products and are currently shared with stakeholders in the implementation of the Marine Strategy Framework Directive (MSFD) such as EEA, EU DG Environment, JRC, and Regional Sea Conventions. Activities are underway for a data exchange with CMEMS. The EMODnet Chemistry data collections include CDI metadata and ODV data, but have a higher quality and coherence as the individual CDI - ODV entries. Developments are underway for establishing an API and GUI for facilitating sub-setting and retrieval of these data collections. Once operational, this service can provide an additional channel to be added to the Blue-Cloud data discovery and access service.

#### **3.3.2.4 Discovery and access mechanisms - how does it function**

See description for SeaDataNet CDI data discovery and access service

#### **3.3.2.5 Metadata format(s) - short overview and references to detailed documentation**

See description for SeaDataNet CDI data discovery and access service

#### **3.3.2.6 Data format(s) - short overview and references to detailed documentation**

See description for SeaDataNet CDI data discovery and access service. Use is made of SeaDataNet ODV4 ASCII and SeaDataNet NetCDF (CF) formats and the specific formats for beach and seafloor litter.

#### **3.3.2.7 Use of controlled vocabularies - which, where, how**

See description for SeaDataNet CDI data discovery and access service.

#### **3.3.2.8 Data access policy - if yes, which and how deployed**

See description for SeaDataNet CDI data discovery and access service.

#### **3.3.2.9 Any web services and API's - URLs, function, how to operate**

See description for SeaDataNet CDI data discovery and access service.

#### **3.3.2.10 Any suggestions for aggregating data sets as collections in order to scale down number of too many entries and serve users with distinctive metadata entries as part of the planned Blue-Cloud data discovery and access service**

See description for SeaDataNet CDI data discovery and access service.

#### **3.3.2.11 Hosting environment**

See description for SeaDataNet CDI data discovery and access service

#### **3.3.2.12 Organisational aspects (main operator(s); data providers)**

See description for SeaDataNet CDI data discovery and access service. Data Providers: currently >60 data centres consisting of NODCs, marine research institutes, and marine environmental monitoring agencies.

#### **3.3.2.13 Contact details**

CDI main operator contact: MARIS, The Netherlands, Dick M.A. Schaap – dick@maris.nl

#### **3.3.2.14 Conclusion for EMODnet Chemistry**

EMODnet Chemistry makes use of the SeaDataNet CDI data discovery and access service. Therefore, no separate solutions need to be developed for Blue-Cloud.

The aggregated, harmonized and validated data collections for eutrophication, contamination, acidification and marine litter, as regularly produced by EMODnet Chemistry, are also relevant for Blue-Cloud purposes. Developments are underway for establishing an API and GUI for facilitating sub-setting and retrieval of these data collections. Once operational, this service will provide an additional channel to be added to the Blue-Cloud data discovery and access service.

### **3.4 EuroArgo - Argo**

The EuroArgo ERIC allows active coordination and strengthening of the European contribution to the international Argo program. Its main objectives are to provide, deploy and operate the European contribution to the global array of Argo floats (currently around 800 floats, ¼ of the global array) and an enhanced coverage of European seas, to expand towards biogeochemistry, greater depths and high latitudes and to provide access to quality-controlled data and derived products.

The EuroArgo ERIC also provides access to quality-controlled data and derived products. This is done by Ifremer in France by hosting one of the two Global Data Assembly Centres (GDACs). The other Argo GDAC is hosted by FNMOC (the Fleet Numerical Meteorology and Oceanography Centre) in USA. The Argo GDACs assemble all data observed by the global array and distribute them worldwide both in real time and in delayed mode. This implicates that the EuroArgo GDAC at Ifremer gives access to the global Argo dataset. However, additional FAIR services are being developed on the Ifremer Euro-Argo GDAC within the ENVRI-FAIR and Euro-Argo-RISE projects.

#### **Core-Argo array**

The broad-scale global array of temperature/salinity profiling floats, known as Argo, has already grown to be a major component of the ocean observing system. Argo is a standard which is an example for other developing ocean observing systems. Argo provides good examples on various topics such as how to collaborate internationally, how to develop a data management system and

how to change the way scientists think about collecting data. Argo float deployments began in 2000 and currently there are circa 4000 Argo floats active.

### BGC-Argo array

Biogeochemical-Argo aims at developing a global network of biogeochemical sensors on Argo profiling floats. The concept of global robotic biogeochemical measurements was articulated in a Community White Paper (Gruber et al., 2007) that was supported by the International Ocean Carbon Coordinating Project (IOCCP) and the US Ocean Carbon and Biogeochemistry Program (US-OCB). This was followed by a Scoping Workshop funded by the US Ocean Carbon and Biogeochemistry Program (Johnson et al., 2009) and an International Ocean Color Coordinating Group (IOCCG) supported working group (IOCCG, 2011). Extensive discussions were held at the OceanObs 09 meeting and were subsequently reported into two community White Papers (Gruber et al., 2010; Claustre et al., 2010). Target for the global array is to have 1000 fully equipped BGC-Argo active floats with a uniform spatial distribution. Euro-Argo aims at contributing to ¼ of the global effort, which represents 250 active BGC floats. These will collect next to the regular Temperature, Salinity and Depth the following BGC parameters: Oxygen concentration; Nitrate concentration; pH; Chlorophyll a concentration; Suspended particles; and Downwelling irradiance.

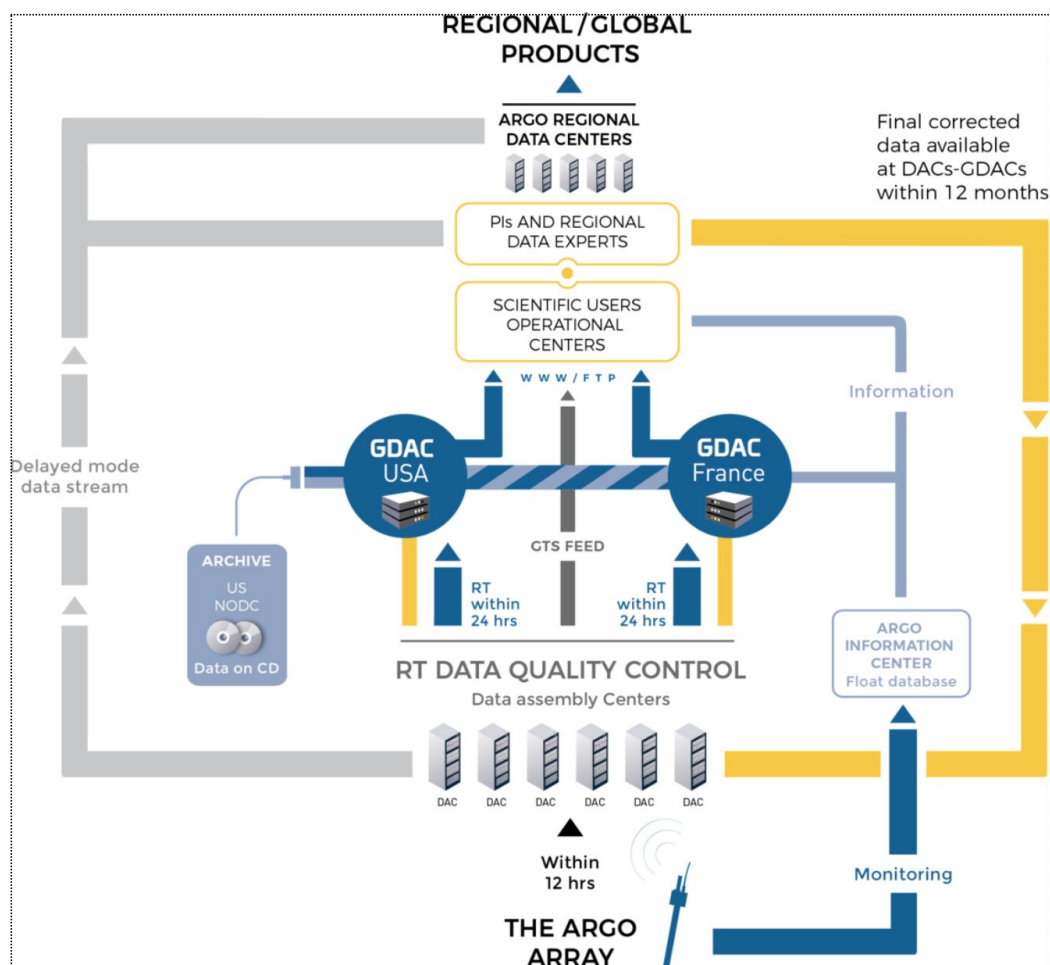


Image 3.4.1: Overview of Argo organisation



### 3.4.1 Data discovery and access service component

The EuroArgo portal features a dashboard which provides a facet search including dynamic map for discovery of Argo floats and open access to its data sets. Also, it is possible to retrieve the whole Argo data collection by a DOI and associated landing page with descriptive metadata about the collection.

Note: In the framework of ENVRI-FAIR and EOSC-hub, a development is underway for an additional data discovery and access service. This will be described in paragraph 3.4.1.5.

#### 3.4.1.1 Name

Argo floats dashboard

Argo DOI landing page

#### 3.4.1.2 Web address

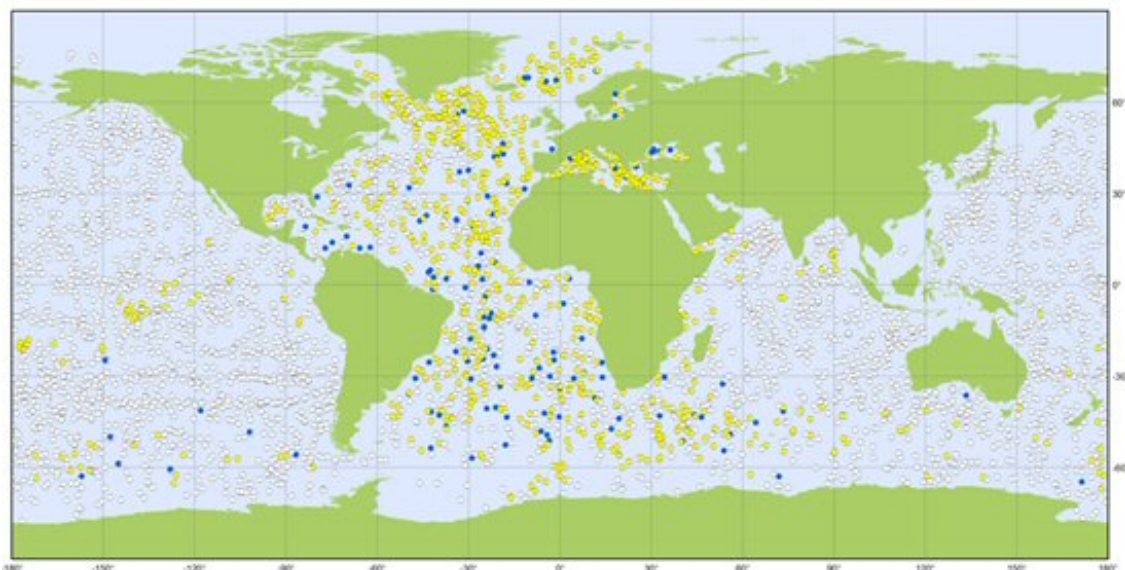
<https://fleetmonitoring.euro-argo.eu/dashboard>

<https://doi.org/10.17882/42182>

#### 3.4.1.3 Types and number of data sets and/or data products

Argo collects salinity/temperature and biogeochemical profiles from an array of robotic floats that populate the ice-free oceans that are deeper than about 2000m. They also give information on the surface and subsurface currents. Most profiles are made up of about 200 data points, but floats with high speed communications may be sending many more data points given the higher bandwidth. In total there are currently 15.400 Argo floats which generated more than 2 million files. Metadata and data for profiles and trajectories, including technical info are made available as NetCDF (CF) files.

In addition, Argo products are generated and made available as gridded fields. See: [http://www.argo.ucsd.edu/Gridded\\_fields.html](http://www.argo.ucsd.edu/Gridded_fields.html)



*Image 3.4.2: Overview map of EuroArgo floats – January 2020*

### 3.4.1.4 Discovery and access mechanisms - how does it function

For discovery and access, there are a number of mechanisms:

- GDAC interactive data selection  
<http://www.argodatamgt.org/Access-to-data/Argo-data-selection>
- GDAC floats dashboard  
<https://fleetmonitoring.euro-argo.eu/dashboard>
- GDAC Thredds server API  
<http://tds0.ifremer.fr/thredds/catalog/CORIOLIS-ARGO-GDAC-OBS/catalog.html>
- GDAC ERDDAP data server API  
<http://www.ifremer.fr/erddap/tabledap/ArgoFloats.graph>

Direct access mechanisms

- GDAC ftp servers  
<ftp://ftp.ifremer.fr/ifremer/argo>
- GDAC DOI (Data Object Identifiers)  
<http://www.argodatamgt.org/Access-to-data/Argo-DOI-Digital-Object-Identifier>
- GDAC synchronization service (rsync)  
<http://www.argodatamgt.org/Access-to-data/Argo-GDAC-synchronization-service>  
<https://doi.org/10.17882/42182>

The dashboard provides a facet search including dynamic map for discovery of Argo floats and open access to its data sets.

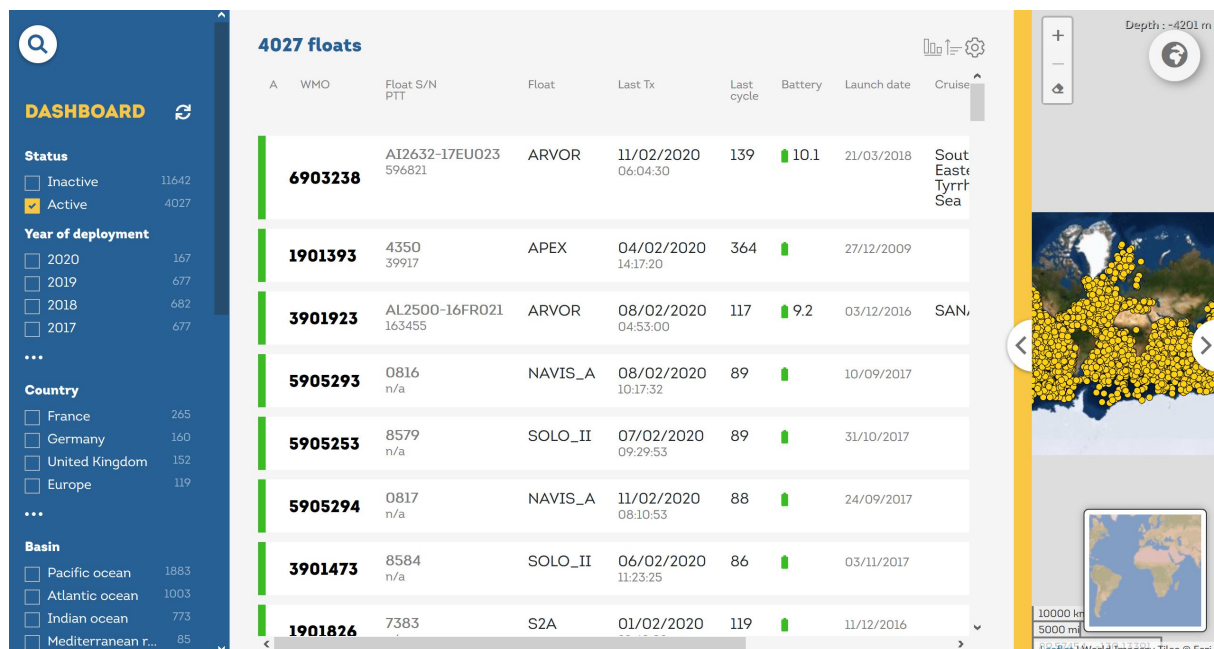


Image 3.4.3: Facet search of Argo floats via dashboard

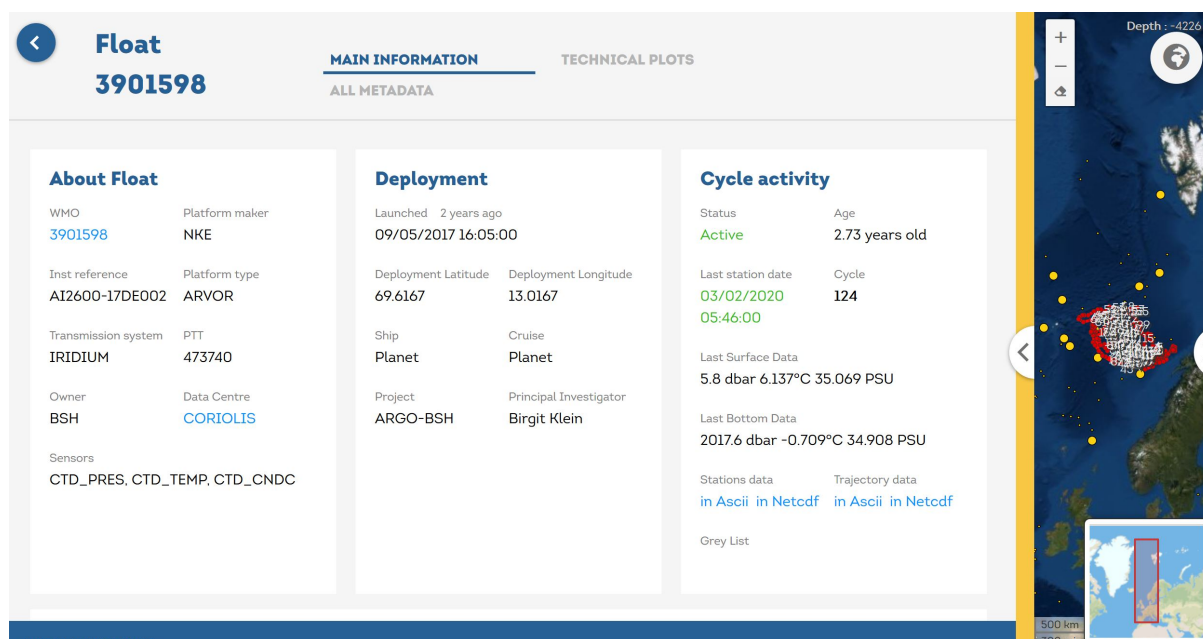


Image 3.4.4: Argo detailed page in dashboard with data access options

### 3.4.1.5 New data discovery and access service developments as part of ENVRI-FAIR, EA-RISE and EOSC-HUB projects

The Argo data are well findable; however, a FAIRness assessment in ENVRI-FAIR concluded that a rich and efficient search service at data level is missing. Therefore Euro-Argo is underway with developing such an advanced service as part of ENVRI-FAIR, EA-RISE and EOSC-hub projects.

#### ENVRI-FAIR, Euro-Argo RI implementation plan

19/09/2019

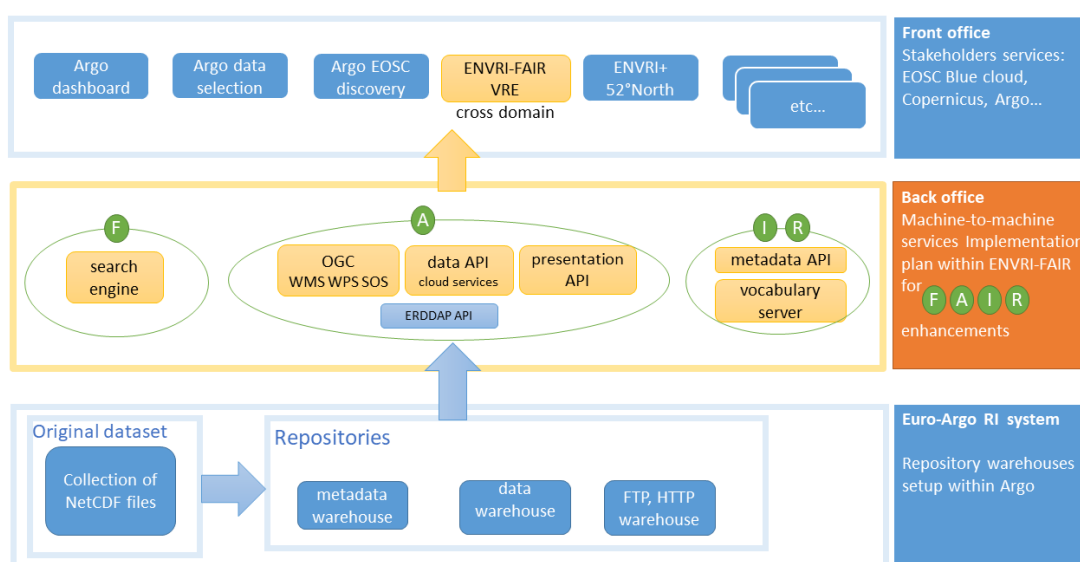


Image 3.4.5: Euro-Argo implementation plan for becoming more FAIR

The approach is to develop API cloud services on top of the collection of Argo NetCDF files. The API services should enable data discovery, visualization, download, and subscription with rich sub-setting capabilities. **Findability** should be improved by implementing a search engine service with possibly OpenSearch on top of an Elasticsearch metadata repository. This is planned for release around end Q1 2021. **Accessibility** should be improved for machine-to-machine interaction by implementing a metadata API, possibly with Elasticsearch or SOL-R on metadata repository (end Q1 2021), and a data API, possibly with Cassandra, Parquet or HBASE on data repository (end Q1 2021). The improvement also includes wider adoption of SeaDataNet vocabularies for improving **interoperability**.

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

Each Argo float has a NetCDF (CF) metadata file with detailed information such as float type, serial number, scientist in charge of the float, scientist in charge of delayed mode adjustments, sensors, sensors serial numbers, mission, etc. The Argo metadata format is documented in the Argo user's manual: <https://doi.org/10.13155/29825>

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

Each Argo float has a NetCDF (CF) profiles and trajectory file with detailed information such as float type, serial number, scientist in charge of the float, scientist in charge of delayed mode adjustments, sensors, calibration details, sensor adjustments, physical parameters cf standard names, error on measurements, adjusted values, etc. The used format is NetCDF CF V3.1. The Argo profiles and trajectory formats are documented in the Argo user's manual: <https://doi.org/10.13155/29825>

### 3.4.1.8 Use of controlled vocabularies - which, where, how

Argo controlled vocabulary is documented in the Argo user's manual: <https://doi.org/10.13155/29825>

Argo physical parameters are linked and available from SeaDataNet vocabulary server P01 and P06 vocabularies. Example for Practical salinity parameter vocabulary:

parameter name	long_name	cf_standard_name	unit	sdn_parameter_urn	sdn_parameter_uri	sdn_uom_urn	sdn_uom_uri
PSAL	Practical salinity	sea_water_salinity	psu	SDN:P01::PSALST01	<a href="http://vocab.nerc.ac.uk/collection/P01/current/PSALST01/">http://vocab.nerc.ac.uk/collection/P01/current/PSALST01/</a>	SDN:P061::UUUUU	<a href="http://vocab.nerc.ac.uk/collection/P06/current/UUUUU/">http://vocab.nerc.ac.uk/collection/P06/current/UUUUU/</a>

### 3.4.1.9 Data access policy - if yes, which and how deployed

Open and free access, [CC-BY](#) licence.

### 3.4.1.10 Any web services and API's - URLs, function, how to operate

The following web services are provided:

- GDAC synchronization service (rsync)  
<http://www.argodatamgt.org/Access-to-data/Argo-GDAC-synchronization-service>  
<https://doi.org/10.17882/42182>
- GDAC Thredds server API  
<http://tds0.ifremer.fr/thredds/catalog/CORIOLIS-ARGO-GDAC-OBS/catalog.html>

- GDAC ERDDAP data server API  
<http://www.ifremer.fr/erddap/tabledap/ArgoFloats.graph>

#### **3.4.1.11 Any suggestions for aggregating data sets as collections in order to scale down number of too many entries and serve users with distinctive metadata entries as part of the planned Blue-Cloud data discovery and access service**

Developments are underway for an advanced discovery and access interface by means of indexing metadata with ElasticSearch and including data in Cassandra dbms for interactive queries and Parquet for Spark highly parallel subsetting. It is recommended to make use of this new advanced interface for serving the Blue-Cloud data discovery and access service.

#### **3.4.1.12 Hosting environment**

The EuroArgo GDAC is hosted at Ifremer in France by means of the Coriolis infrastructure.

#### **3.4.1.13 Organisational aspects (main operator(s); data providers)**

Worldwide, there are 11 Data Assembly Centers which process their national/regional floats data (NetCDF formatting, quality control). Those data sets are then aggregated in parallel by the GDACs in USA and France. The GDACs distribute data to the user community.

#### **3.4.1.14 Contact details**

EuroArgo main operator contact at Ifremer: [euroargo@ifremer.fr](mailto:euroargo@ifremer.fr)

#### **3.4.1.15 Conclusions for EuroArgo – Argo**

EuroArgo operates a number of web services for discovery and access to the ArgoFloat data sets. These can be used for the first release of the Blue-Cloud data discovery and access service.

EuroArgo is developing advanced services as part of the ENVRI-FAIR, EIOSC-hub, and EA-RISE projects, which should be followed closely as near-future candidate for coupling to the Blue-Cloud.

## **3.5 EurOBIS – EMODnet Biology**

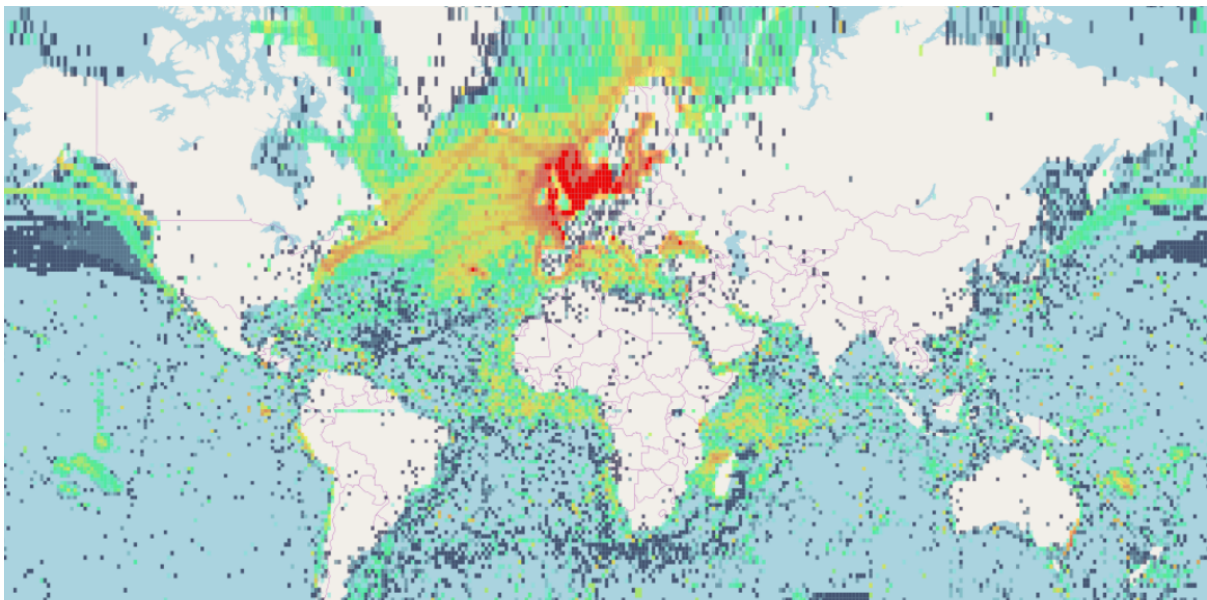
EurOBIS ([www.eurobis.org](http://www.eurobis.org)) was developed by the Flanders Marine Institute in 2004, within the framework of the MarBEF project (MARine Biodiversity and Ecosystem Functioning). It brings together biogeographic data collected within European marine waters, or by European researchers and institutes outside Europe. It focuses on taxonomy and distribution records in space and time and offers a number of online tools to easily query and visualise the data. Currently, EurOBIS holds 889 datasets, representing 62,299 species and 24 million distribution records.

With more than 6 million distribution records, fish are the most common in the database, followed by (sea) birds and marine mammals. At a species level, Atlantic herring, dab, whiting and Atlantic cod take the lead with 650-780,000 distribution records each, with some of them going back to the early 17th century, predating Linnaeus and Darwin. The oldest record is for a masked crab, registered in 1507 in UK waters and added to the EurOBIS database in May 2019. Over the years, the EurOBIS database structure has evolved, making it possible to not only capture presence or abundance of species, but also e.g. biomass data and length measurements in a



standardised and structured way. Remarkably, even ‘blubber thickness in marine mammals’ is part of the EurOBIS information system, with 247 measurements made on the beach, in species such as harbour porpoise, grey seal, harbour seal, white-beaked dolphin or bottlenose dolphin.

Similar to what happens with other regional nodes, EurOBIS data flow to the global initiative Ocean Biogeographic Information System ([OBIS](#)) and eventually become available via the Global Biodiversity Information Facility ([GBIF](#)), hosting global marine and terrestrial distribution data. From 2009 EurOBIS became the backbone of the European Marine Observation and Data Network Biology ([EMODnet Biology](#)), allowing a flow of EurOBIS data through its portal. In 2014 EurOBIS became part of the central [Species Information Backbone of LifeWatch](#), which aims at standardizing species data and integrating the distributed biodiversity data and taxonomic repositories and operating facilities as well as filling the gaps in our knowledge. The EurOBIS data management team is supported by [LifeWatch Belgium](#), part of the European LifeWatch E-Science Infrastructure for Biodiversity and Ecosystem Research.



*Image 3.5.1: Map interface with overview of EurOBIS observations*

### **3.5.1 Data discovery and access service component**

The data can be discovered and accessed by means of the [Data Catalog](#) or the [Data Download Toolbox](#).

#### **3.5.1.1 Name**

The data discovery and access services are available from the EMODnet Biology website.

#### **3.5.1.2 Web address**

Data Catalog: <https://www.emodnet-biology.eu/data-catalog>

Data Download Toolbox:

<https://www.emodnet-biology.eu/toolbox/en/download/occurrence/explore>

### 3.5.1.3 Types and number of data sets and/or data products

Currently there are 896 datasets, see [http://www.eurobis.org/dataset\\_list](http://www.eurobis.org/dataset_list)

### 3.5.1.4 Discovery and access mechanisms - how does it function

Download toolbox: stepwise selection and filtering, see tutorial:

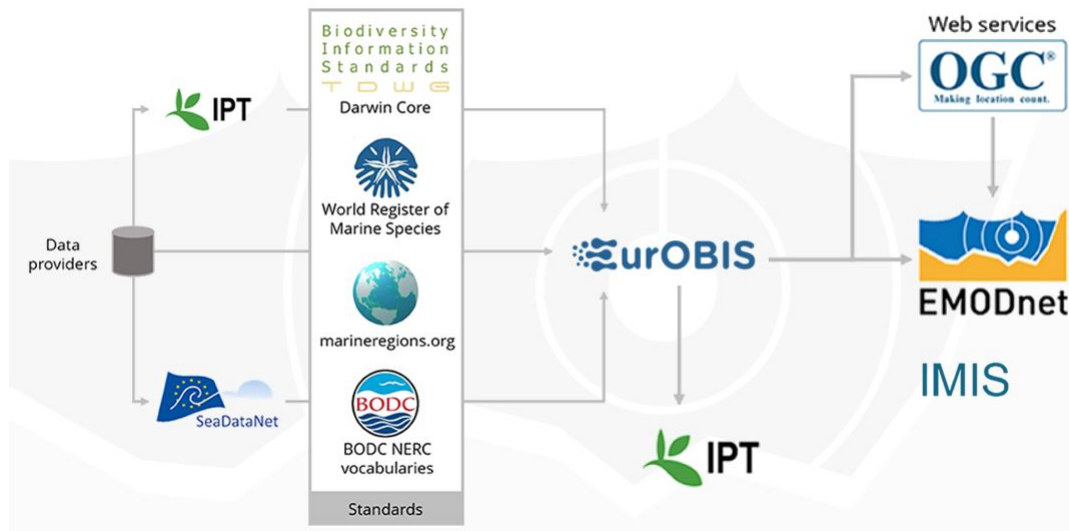
<https://www.emodnet-biology.eu/tutorials>

Data Catalog with search on different metadata fields:

<https://www.emodnet-biology.eu/data-catalog>

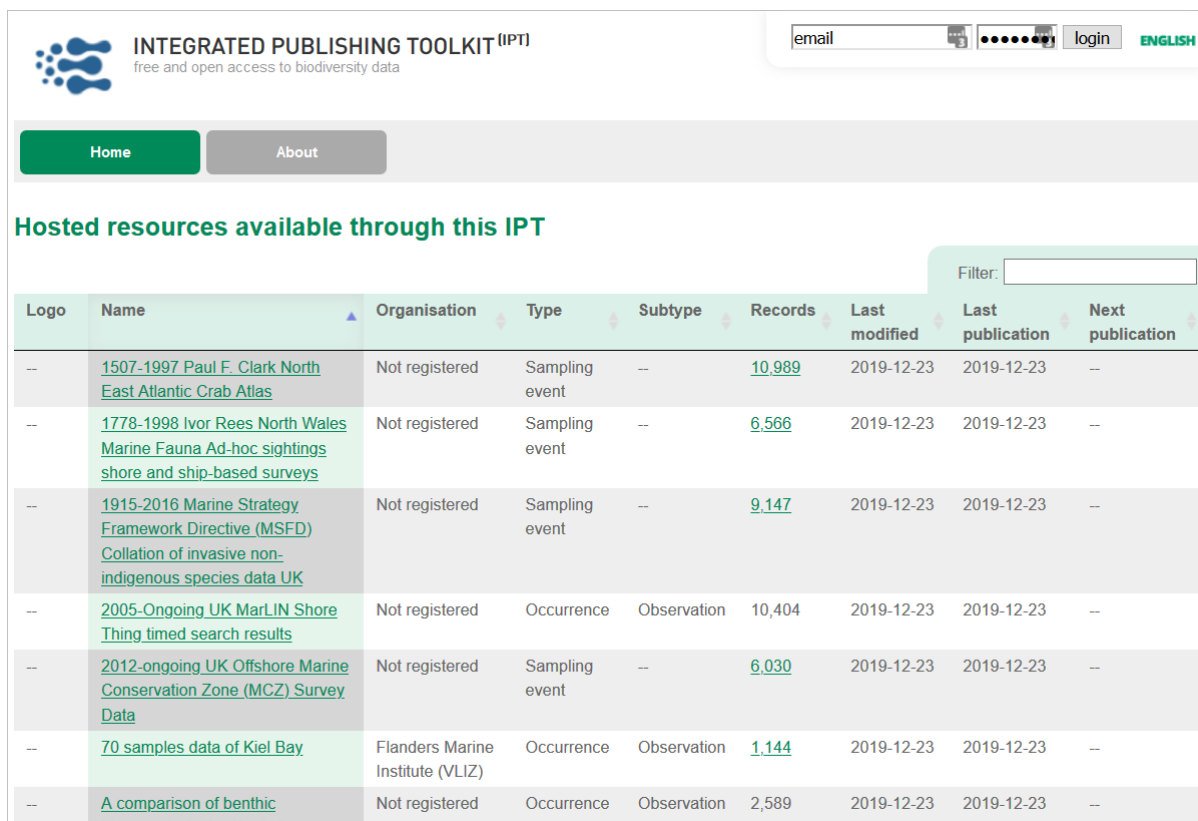
Geonetwork catalogue: <http://www.emodnet.eu/geonetwork>

The architecture of the EurOBIS – EMODnet Biology system is illustrated below:



*Image 3.5.2: EurOBIS – EMODnet Biology system architecture and metadata – data flow*

The IPT stands for Integrated Publishing Toolkit. This could be very suitable as catalogue / discovery service endpoint for the Blue-Cloud data discovery and access service: [ipt.vliz.be/eurobis/](http://ipt.vliz.be/eurobis/)



**INTEGRATED PUBLISHING TOOLKIT (IPT)**  
free and open access to biodiversity data

email   login **ENGLISH**

**Home** **About**

**Hosted resources available through this IPT**

Filter:

Logo	Name	Organisation	Type	Subtype	Records	Last modified	Last publication	Next publication
--	<a href="#">1507-1997 Paul F. Clark North East Atlantic Crab Atlas</a>	Not registered	Sampling event	--	<a href="#">10,989</a>	2019-12-23	2019-12-23	--
--	<a href="#">1778-1998 Ivor Rees North Wales Marine Fauna Ad-hoc sightings shore and ship-based surveys</a>	Not registered	Sampling event	--	<a href="#">6,566</a>	2019-12-23	2019-12-23	--
--	<a href="#">1915-2016 Marine Strategy Framework Directive (MSFD) Collation of invasive non-indigenous species data UK</a>	Not registered	Sampling event	--	<a href="#">9,147</a>	2019-12-23	2019-12-23	--
--	<a href="#">2005-Ongoing UK MarLIN Shore Thing timed search results</a>	Not registered	Occurrence	Observation	10,404	2019-12-23	2019-12-23	--
--	<a href="#">2012-ongoing UK Offshore Marine Conservation Zone (MCZ) Survey Data</a>	Not registered	Sampling event	--	<a href="#">6,030</a>	2019-12-23	2019-12-23	--
--	<a href="#">70 samples data of Kiel Bay</a>	Flanders Marine Institute (VLIZ)	Occurrence	Observation	<a href="#">1,144</a>	2019-12-23	2019-12-23	--
--	<a href="#">A comparison of benthic biodiversity in the North Sea</a>	Not registered	Occurrence	Observation	2,589	2019-12-23	2019-12-23	--

Image 3.5.3: EurOBIS – EMODnet Biology IPT service

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

All EurOBIS/EMODnet Biology datasets are described in the [EMODnet Biology Catalogue](#). The EMODnet Catalogue is part of a larger integrated metadata system called the Integrated Marine Information System ([IMIS](#)). As IMIS is hosted by Flanders Marine Institute (VLIZ), it is focused on Flanders and supplemented by the scientific output of projects involving VLIZ. It contains metadata about all people, institutes, publications, projects and datasets that are about or involved in marine science and links these different modules together. It should be noted that all these metadata records are interlinked, which allows easy discovery of the scientific output of an institute or a scientist.

IMIS is INSPIRE and ISO19115 compliant, and can export in JSON, EML & XML:

<http://www.vliz.be/imis?page=webservices>

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

EurOBIS makes use of the Darwin Core model. The conceptual data model of the Darwin Core Archive is a “**star schema**” (Robertson et al. 2014):

- **Core record**, such as an occurrence or an event, as the center of the star.
- **Extension records**, radiating out of the star, can optionally be associated with the core, linked by database keys such as an ID column.

The Darwin Core Event Core is illustrated below:

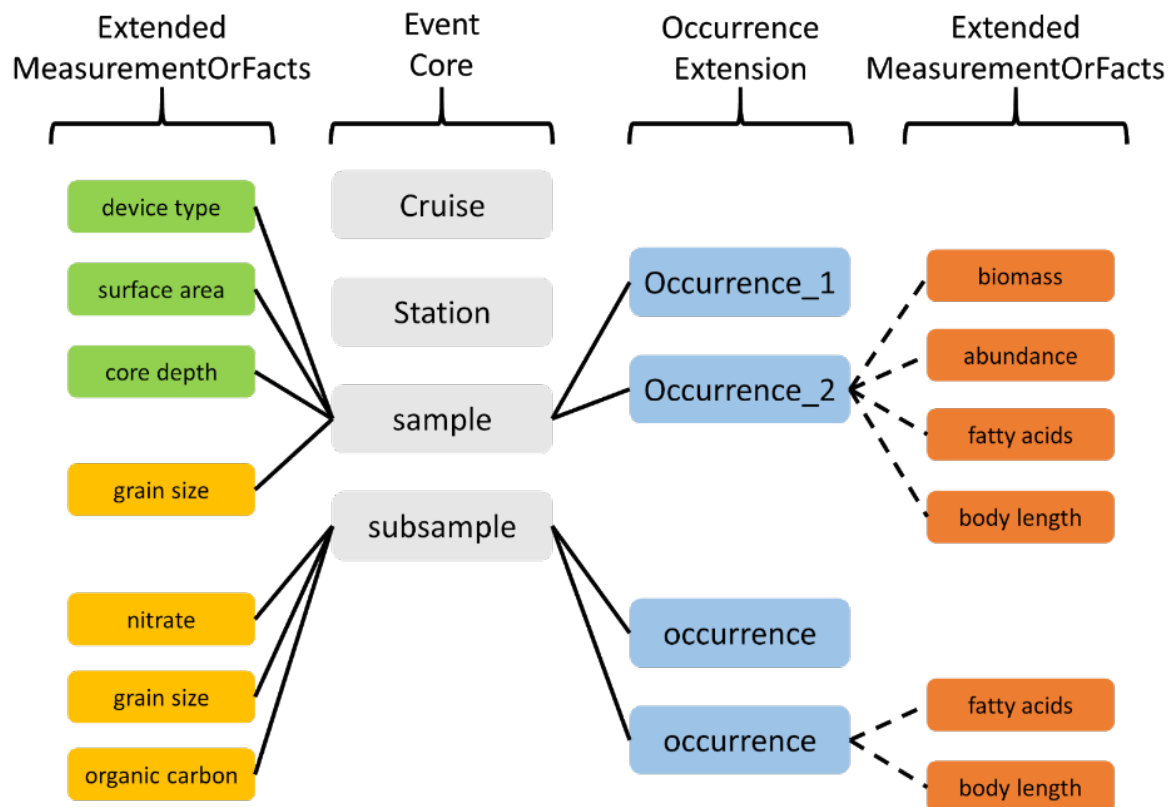
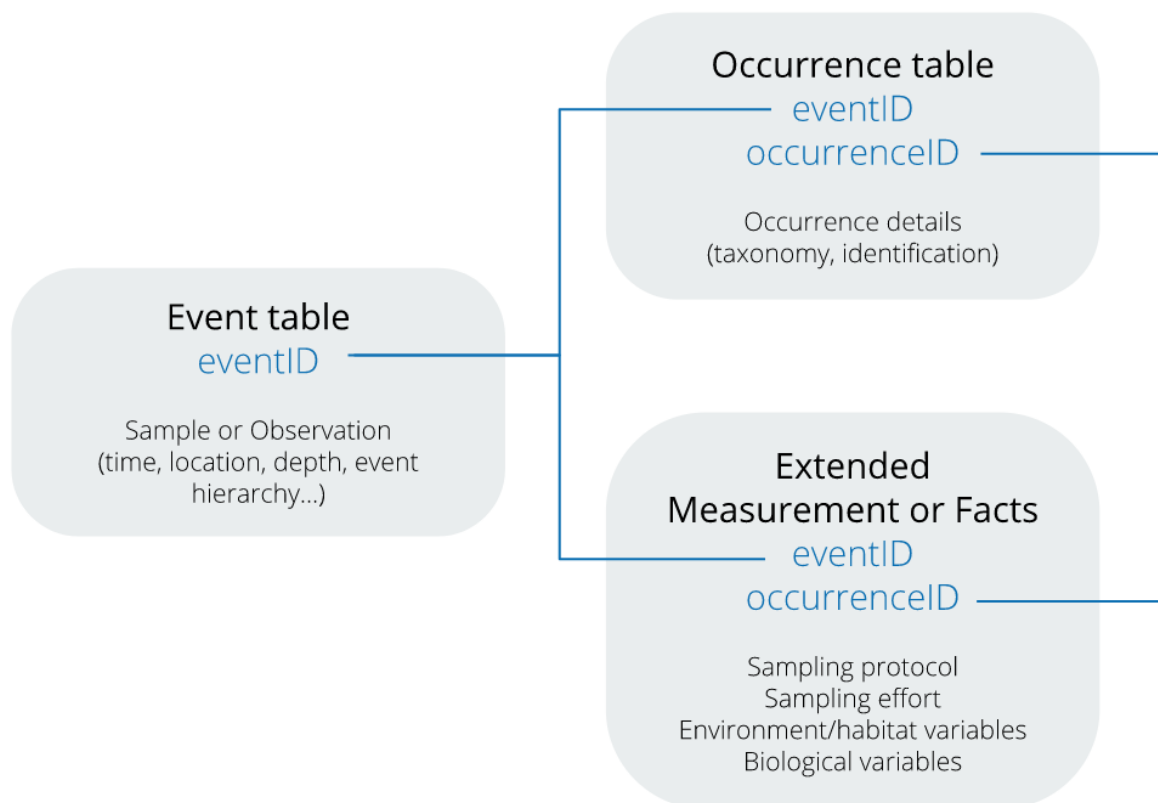


Image 3.5.4: Darwin Core Event Core

OBIS-Env is an extension which facilitates including environmental data which were observed in conjunction and together with the biological observations.

The [DwC terms](#) that are most relevant to EMODnet Biology format are the following (those in **bold** are mandatory):

- Event table
  - datasetName, **eventID**, parentEventID, **eventDate**, **institutionCode**, habitat, type, minimumDepthInMeters, maximumDepthInMeters, **decimalLatitude**, **decimalLongitude**, coordinateUncertaintyInMeters, footprintWKT, modified
- Occurrence table
  - **eventID**, **occurrenceID**, **scientificName**, scientificNameAuthorship, **scientificNameID**, kingdom, taxonRank, identificationQualifier, **occurrenceStatus**, **basisOfRecord**, modified
- Extended MeasurementorFact table
  - measurementID, **eventID**, occurrenceID, **measurementType**, **measurementTypeID**, **measurementValue**, measurementValueID, measurementUnit, measurementUnitID, measurementAccuracy, measurementRemarks



*Image 3.5.5: OBIS-ENV model consisting of DwC Event core + eMoF extension*

**Reference:**

De Pooter D, Appeltans W, Bailly N, Bristol S, Deneudt K, Eliezer M, Fujioka E, Giorgetti A, Goldstein P, Lewis M, Lipizer M, Mackay K, Marin M, Moncoiffé G, Nikolopoulou S, Provoost P, Rauch S, Roubicek A, Torres C, van de Putte A, Vandepitte L, Vanhoorne B, Vinci M, Wambiji N, Watts D, Klein Salas E, Hernandez F (2017) Toward a new data standard for combined marine biological and environmental datasets - expanding OBIS beyond species occurrences. Biodiversity Data Journal 5: e10989. <https://doi.org/10.3897/BDJ.5.e10989>

**More info:**

[http://www.eurobis.org/data\\_formats](http://www.eurobis.org/data_formats)

<https://www.emodnet-biology.eu/tutorials>

<https://classroom.oceanteacher.org/course/view.php?id=328>

### 3.5.1.7 Use of controlled vocabularies - which, where, how

Datasets are thoroughly checked before being ingested in the EuroBIS /EMODnet Biology database. The datasets have to fulfill different criteria and use controlled vocabularies before they will be ingested. See previous section for references.

**Vocabularies:** NERC Vocabulary Server



**Taxonomy:** World Register of Marine Species LSID

**Geography:** MarineRegions.org

See <http://www.eurobis.org/standards>

### **3.5.1.8 Data access policy - if yes, which and how deployed**

Download toolbox: download interface with necessary fields: Country, Data purpose

Webservices: none

ca. 70% CC BY or less strict, ca. 30% CC BY-NC or CC BY-NC-SA

### **3.5.1.9 Any web services and API's - URLs, function, how to operate**

WFS: <http://geo.vliz.be/geoserver/Emodnetbio/wfs>

Examples and explanation see <https://www.emodnet-biology.eu/emodnet-biology-api>

Tutorials in R are available at: <http://www.opensealab.eu/data2019> > R Tutorials > EMODnet > EMODnet Biology

### **3.5.1.10 Any suggestions for aggregating data sets as collections in order to scale down number of too many entries and serve users with distinctive metadata entries as part of the planned Blue-Cloud data discovery and access service**

EurOBIS has already aggregated its data to data collections level, which should be perfect for the Blue-Cloud data discovery and access service.

### **3.5.1.11 Hosting environment**

PostgreSQL+PostGIS database, with web services that are provided through GeoServer on premise, at the Flanders Marine Institute.

### **3.5.1.12 Organisational aspects (main operator(s); data providers)**

Main operator: VLIZ Data Centre + VLIZ IT

Data providers: multiple European partners

### **3.5.1.13 Contact details**

[bio@emodnet.eu](mailto:bio@emodnet.eu)

[info@eurobis.org](mailto:info@eurobis.org)

### **3.5.1.14 Conclusions EurOBIS – EMODnet Biology**

EurOBIS – EMODnet Biology operates a number of web services for discovery and access to the EurOBIS data sets. Of these, the endpoint of the Integrated Publishing Toolkit (IPT) seems to be most suited for connecting to the Blue-Cloud data discovery and access service.

## 3.6 EcoTaxa

EcoTaxa is a web application dedicated to the visual exploration and the taxonomic annotation of images that illustrate the beauty of planktonic biodiversity. EcoTaxa was born from the experience developed at Laboratoire d'Océanographie de Villefranche (LOV) regarding the quantitative, high-throughput imaging of plankton and of the Oceanomics project which covered the exploitation of data collected during the Tara Oceans cruise, including quantitative imaging. It is now developed mainly through the WWWPIC project funded by the Belmont Forum and as part of the Blue-Cloud project. The aim of EcoTaxa is to centralize images of plankton, to allow their collaborative sorting along a universal taxonomy and to accelerate it through machine learning. It produces ecological data in the form of concentration and biovolume of organisms in a given taxon, at a given station (lat, lon, time).

Visitors have free access to the specimens that have been already identified by taxonomist experts. They can explore the database by navigating along the UniEuk taxonomic tree which aims at unifying taxonomic names and tree according to reliable and curated molecular phylogenies. It encompasses the whole Eukaryotic and Prokaryotic lineages (Viruses coming soon) that have been molecularly described. Then images can be filtered according to several sample criteria. Tools are provided to support the annotation of large image datasets by supervised machine learning prediction.

### 3.6.1 Data discovery and access service component

Currently, data discovery is manual. As part of the WWWPIC and Blue-Cloud projects it is aimed to build an API to allow programmatic browsing of the datasets and download of the data.

#### 3.6.1.1 Name

EcoTaxa

#### 3.6.1.2 Web address

<http://ecotaxa.obs-vlfr.fr>

#### 3.6.1.3 Types and number of data sets and/or data products

Currently, EcoTaxa contains circa 95 million images of which circa 40 million have been annotated in about 1300 datasets. Of these, circa 20 million images concern living organisms. The growth rate is circa 1 million images per month. Not all of these datasets will be accessible for Blue-Cloud, because this depends on the data policy of data providers while EcoTaxa does not enforce datasets to be public. It also depends on the ability in the Blue-Cloud project to develop a connector towards EuroBioImaging and/or EMODnet Biology. Priority will be given to the datasets needed by demonstrators.

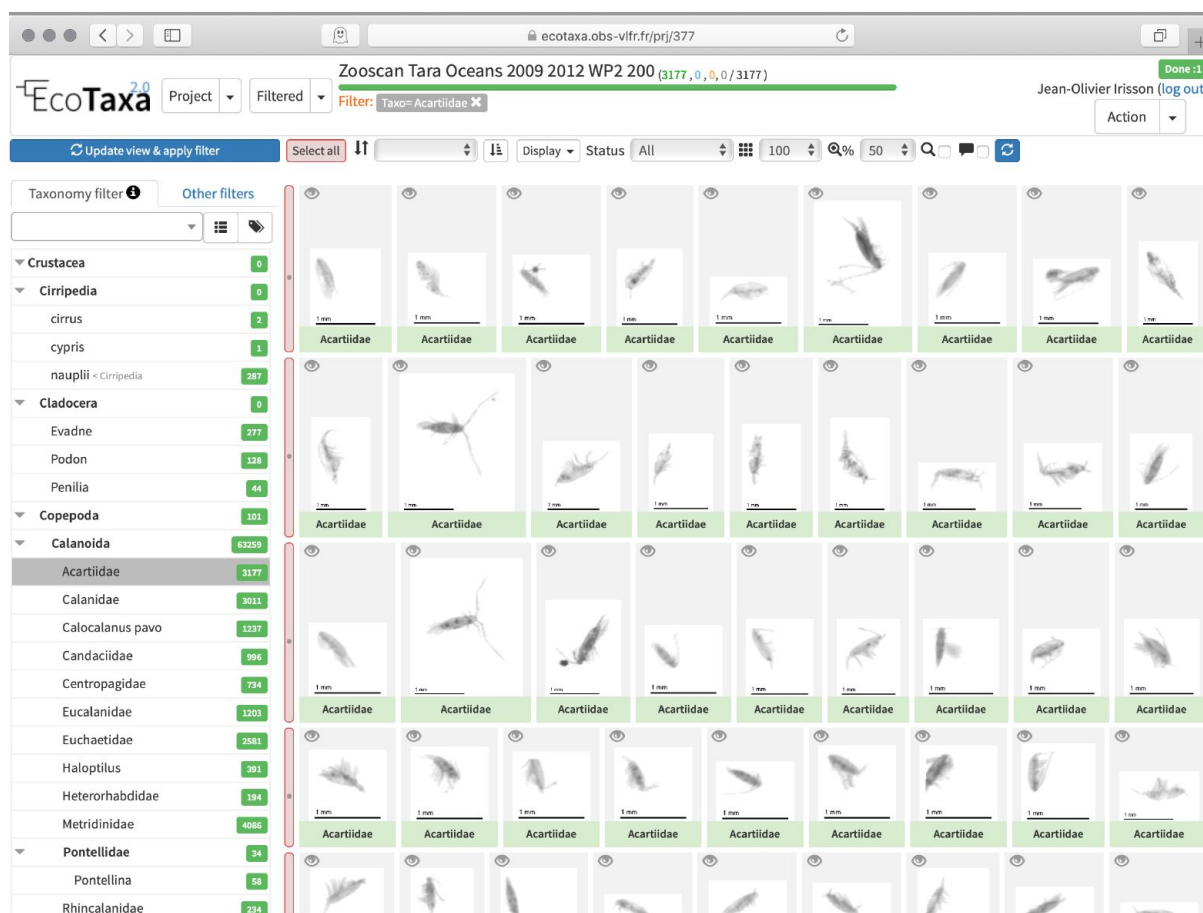
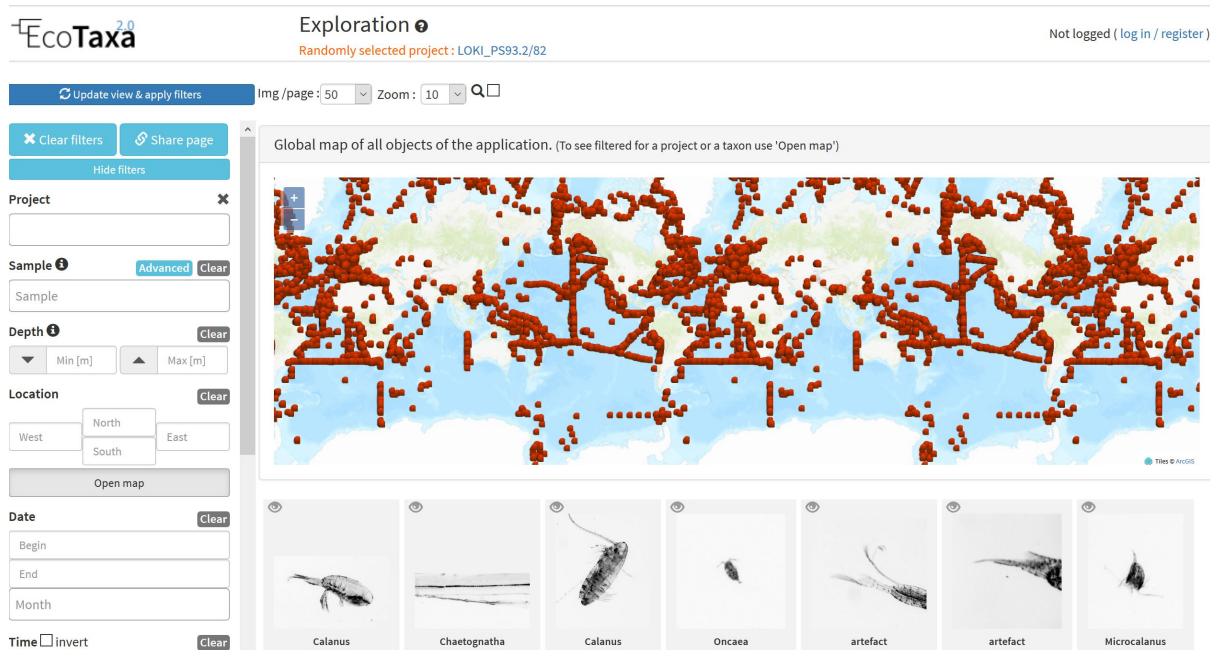


Image 3.6.1: EcoTaxa dashboard for annotation

### 3.6.1.4 Discovery and access mechanisms - how does it function

As indicated earlier, data discovery is currently manual. Validated data from all public datasets can be browsed at <https://ecotaxa.obs-vlfr.fr/explore/> and the datasets themselves are listed at <https://ecotaxa.obs-vlfr.fr/prjothers/>. Data can be downloaded once access to the dataset is granted by the dataset owner.

As part of the WWWPIC and Blue-Cloud projects activities are underway for developing an API which will allow to discover and download data sets by machine-to-machine interfacing. Striving for sustainability, the pursued approach is to make the EcoTaxa database an additional resource of the existing cooperation between the European Marine Biological Research Centre (EMBR) (possibly responsible for operation), Institut Francaise de BioInformatique (IFB) (possibly responsible for hosting), and EuroBioImaging (EMBL-EBI) (possibly responsible for long term archiving). This will secure the long term operation of EcoTaxa, while the API under development will facilitate to integrate the discovery and access of images into EurOBIS – EMODnet Biology from which it will become available for the Blue-Cloud data discovery and access service.



*Image 3.6.2: EcoTaxa interface for exploring images*

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

The metadata fields with homogeneous formats are

- Latitude, longitude (decimal degrees)
- Date, time (ISO8601 format)
- Depth\_min, Depth\_max
- Taxonomic class
- Acquisition instrument

The field names are currently not standardized but will be formatted according to the DarwinCore vocabulary <https://dwc.tdwg.org> as part of the development and planned functioning of the API for discovery and access. Many other metadata fields are stored, but these vary from dataset to dataset to accommodate the differences among imaging instruments and dataset purposes. As part of the API, at least the following other fields will be made homogeneous: concentration, pixel size, object dimensions. These fields are present in almost all datasets.

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

Currently, EcoTaxa exports data in TSV and ODV compatible formats. Activities are underway for standardising more fields, adopting common vocabularies (WoRMS and SeaDataNet) and for adopting the DarwinCore format specification in order to facilitate the planned exchange towards EurOBIS – EMODnet Biology. For that purpose, also an aggregation is foreseen to limit the number of entries. The image below gives more detail how the existing EcoTaxa data format will become more interoperable.

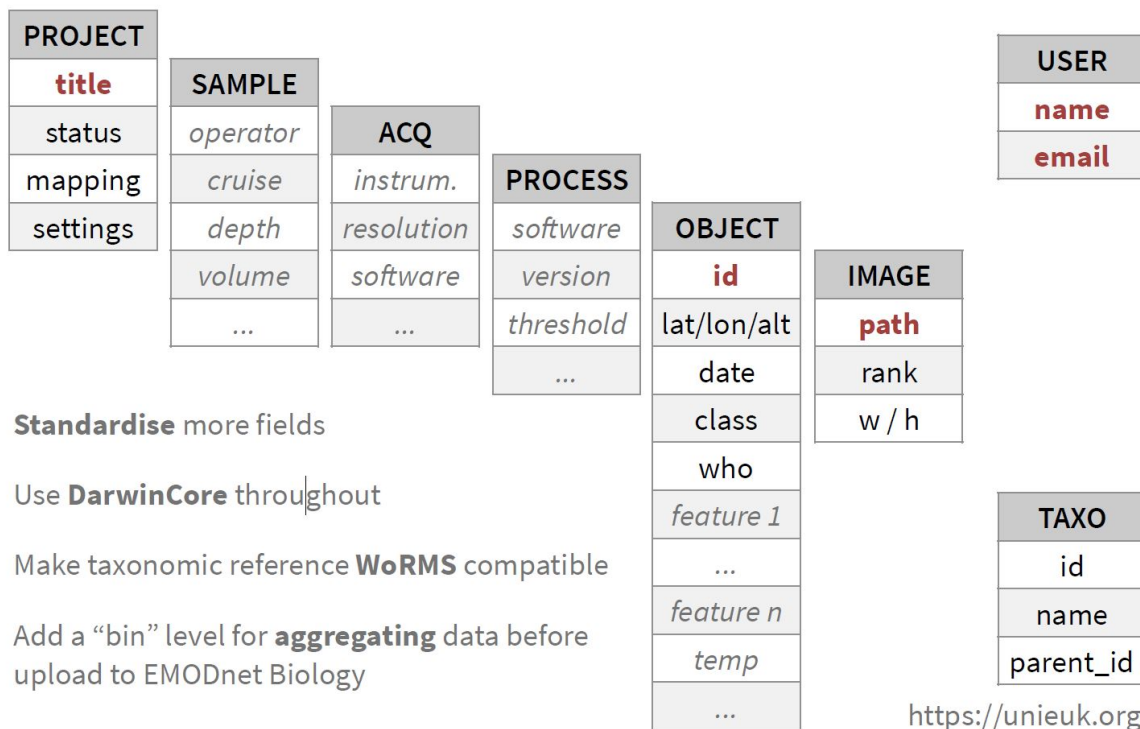


Image 3.6.3: Making the EcoTaxa data model more interoperable for exchange to EurOBIS – EMODnet Biology

### 3.6.1.7 Use of controlled vocabularies - which, where, how

Currently no controlled vocabularies are in use. The adoption and adaptation to DarwinCore will require adopting prevailing vocabularies for every standard field. These are WoRMS (World Register of Marine Species), selected SeaDataNet vocabularies, and MarineRegions.org for geography. See also the description of EurOBIS – EMODnet Biology. Non-standard fields will continue to be defined for each dataset by each dataset manager.

### 3.6.1.8 Data access policy - if yes, which and how deployed

Each dataset is the propriety of its manager(s). They can be made private (i.e. not browsable) or public (i.e. browsable). Data download is only allowed for dataset “members” (the dataset managers are the only ones able to add dataset members).

With the advent of the API, the access rules will be changed: CC licenses will be associated with each dataset (CC-0, CC-BY, CC-BY-NC, C) and data will be directly downloadable from EcoTaxa by anyone if the dataset manager allows.

### 3.6.1.9 Any web services and API's - URLs, function, how to operate

None at the moment. Development of API is underway as part of the WWWPIC and Blue-Cloud projects. See explanations above.



#### **3.6.1.10 Any suggestions for aggregating data sets as collections in order to scale down number of too many entries and serve users with distinctive metadata entries as part of the planned Blue-Cloud data discovery and access service**

Many datasets are comprehensive collections such as all stations of a given scientific cruise. These are considered as units and should stay separate. Several others are pieces of larger datasets (e.g. separate periods within a long-term time series). Those pieces should be joined together to form a single comprehensive unit.

#### **3.6.1.11 Hosting environment**

EcoTaxa is currently hosted at Institut de la Mer de Villefranche (IMEV) which is the parent organisation of LOV. Within 1.5 years it will migrate for hosting to the Institut Français de Bioinformatique (IFB). See also paragraph 3.6.1.4.

#### **3.6.1.12 Organisational aspects (main operator(s); data providers)**

The infrastructure is fully operated by staff at Laboratoire d'Océanographie de Villefranche (LOV) and Institut de la Mer de Villefranche (IMEV) (in Villefranche sur mer, southern France). It is developed by LOV and a subcontractor in France. Data providers are extremely varied; currently, there are 750 registered users from 250 institutions worldwide. The users mostly come from France, Germany, the USA, and to a lesser extent from China, Brasil, Canada, South Africa, and other EU countries. A few users come from yet other countries.

#### **3.6.1.13 Contact details**

LOV - Jean-Olivier Irisson ([irisson@normalesup.org](mailto:irisson@normalesup.org))

#### **3.6.1.14 Conclusion EcoTaxa**

As part of the Blue-Cloud, EcoTaxa metadata – data will be integrated in EurOBIS – EMODnet Biology by API which is under development. The coupling to the Blue-Cloud data discovery and access service will then be provided through EurOBIS – EMODnet Biology.

### **3.7 ELIXIR-ENA**

The European Nucleotide Archive (ENA) provides a comprehensive open record of the world's nucleotide sequencing information and a platform for the management and analysis of sequence and related data. Covering raw sequencing data, sequence assembly information, functional annotation and a host of further data types, content is measured in millions of taxa, hundreds of thousands of sequenced libraries and petabytes of storage. ENA is operated by the EMBL European Bioinformatics Institute (EMBL – EBI). ENA is designated by the ELIXIR infrastructure both as a Core Data Resource, and a Deposition Database. For more background information, see: <https://www.ebi.ac.uk/ena/browser/about>.

ENA's portfolio of services include user support (helpdesk, training), web sites (data submissions, browser with search, explore and download functions), RESTful interfaces (data submissions, data discovery, metadata interrogation) and a host of downloadable utilities to support data submissions

and access. As a founding member of the celebrated International Nucleotide Sequence Database Collaboration (INSDC), ENA drives international standards and best practice in its domain.

### 3.7.1 Data discovery and access service component

The ENA browser (<https://www.ebi.ac.uk/ena/browser>) brings together a set of services via web interfaces, build upon underlying APIs. Of relevance for Blue-Cloud are two services: data discovery (metadata search and retrieval) and data retrieval.

#### 3.7.1.1 Name

ENA Data Discovery and ENA Data Retrieval

#### 3.7.1.2 Web address

ENA Data Discovery: <https://www.ebi.ac.uk/ena/browser/advanced-search>

ENA Data Retrieval: <https://www.ebi.ac.uk/ena/browser/home>

#### 3.7.1.3 Types and number of data sets and/or data products

ENA covers many data types in a number of interlinked database tables. A listing can be found at <https://www.ebi.ac.uk/ena/portal/api/results?dataPortal=ena>

Data type/class	Explanation
resultId	Description
analysis_study	Studies used for nucleotide sequence analyses from reads
analysis	Nucleotide sequence analyses from reads
assembly	Genome assemblies
coding_release	Protein coding sequences (Release)
coding_update	Protein coding sequences (Update)
wgs_set	Genome assembly contig sets (WGS)
tss_set	Transcriptome assembly contig sets (TSA)
environmental	Environmental samples
noncoding_release	Non-coding sequences (Release)
noncoding_update	Non-coding sequences (Update)
noncoding	Non-coding sequences
read_study	Studies used for raw reads
read_experiment	Experiments used for raw reads
read_run	Raw reads
sample	Samples
sequence_release	Nucleotide sequences (Release)
sequence_update	Nucleotide sequences (Update)
sequence	Nucleotide sequences
study	Studies
taxon	Taxonomic classification

Table 3.7.1: data types and database components in ENA system

Details can be found at:

<https://ena-browser-docs.readthedocs.io/en/latest/browser/search/advanced.html>

and in the documentation of the ENA Portal API: <https://www.ebi.ac.uk/ena/portal/api/doc>

In numbers:

- **Growth Rate:** 1 new dataset every 6 minutes
- **Data:**  $2 \times 10^9$  sequences and  $1 \times 10^{16}$  base pairs of read data across  $2 \times 10^6$  taxa
- **Usage:** 2,000 submitters; 10x thousands monthly consumers; 10x millions of monthly hits, many times this globally
- **Support:** 46 tickets per day and in-person training delivered to more than 350 users per annum

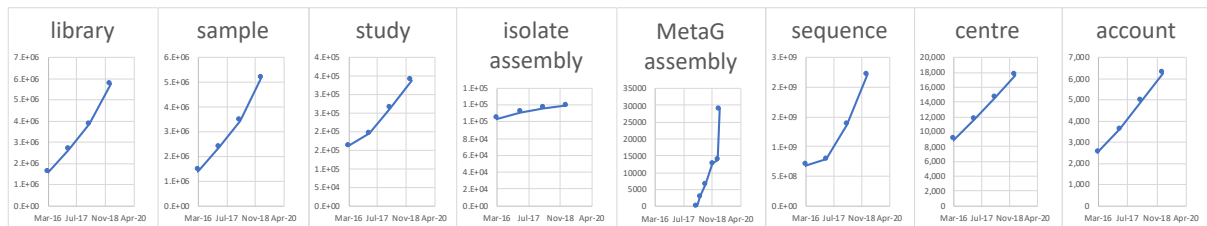


Image 3.7.1: Illustration of growth of selected data types since March 2016

### 3.7.1.4 Discovery and access mechanisms - how does it function

Search queries can be built using UI wizard at <https://www.ebi.ac.uk/ena/browser/advanced-search> and run in the Browser, or using the “Copy Curl Request” button exported as a curl command to be run in the command line. There are many criteria to build a query. A good guidance can be found at:

<https://ena-browser-docs.readthedocs.io/en/latest/browser/search/advanced.html#search-query>

```
curl -X POST -H "Content-Type: application/x-www-form-urlencoded" -d 'result=sample&query=tax_eq(27697)&format=tsv' "https://www.ebi.ac.uk/ena/portal/api/search"
```

Image 3.7.2: Example of ENA query options – in this case by Taxonomy

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

Indexed metadata is available as TSV (Tab Separated Value) by default or as JSON (format=json). Details can be found at:

<https://ena-browser-docs.readthedocs.io/en/latest/browser/search/advanced.html#download-results-report>

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

Data can be retrieved in different formats and with easy file download options through RESTful services:

- EMBL Flatfile format
- FASTA format for sequences
- XML Format

Details about formats:

<https://ena-browser-docs.readthedocs.io/en/latest/browser/search/advanced.html#download-ena-records>

<https://www.ebi.ac.uk/ena/browser/api/xml/SAMEA2619443>

```
<SAMPLE_SET>
  <SAMPLE accession="ERS487982" alias="TARA_X000000488" center_name="Genoscope">
    <IDENTIFIERS>
      <PRIMARY_ID>ERS487982</PRIMARY_ID>
      <EXTERNAL_ID namespace="BioSample">SAMEA2619443</EXTERNAL_ID>
      <SUBMITTER_ID namespace="GSC">TARA_X000000488</SUBMITTER_ID>
    </IDENTIFIERS>
    <TITLE>
      TARA_20090920T1045Z_005_EVENT_PUMP_P_S_(5 m)_PROT_NUC-RNA(100L)_W0.8-5_TARA_X000000488
    </TITLE>
    <SAMPLE_NAME>
      <TAXON_ID>408172</TAXON_ID>
      <SCIENTIFIC_NAME>marine metagenome</SCIENTIFIC_NAME>
    </SAMPLE_NAME>
    <DESCRIPTION>...</DESCRIPTION>
    <SAMPLE_LINKS>
      <SAMPLE_LINK>
        <XREF_LINK>
          <DB>ENA-STUDY</DB>
          <ID>ERP003628,ERP006157,ERP018626</ID>
        </XREF_LINK>
      </SAMPLE_LINK>
      <SAMPLE_LINK>...</SAMPLE_LINK>
      <SAMPLE_LINK>...</SAMPLE_LINK>
      <SAMPLE_LINK>...</SAMPLE_LINK>
      <SAMPLE_LINK>...</SAMPLE_LINK>
      <SAMPLE_LINK>...</SAMPLE_LINK>
    </SAMPLE_LINKS>
    <SAMPLE_ATTRIBUTES>
      <SAMPLE_ATTRIBUTE>
        <TAG>Sampling Campaign</TAG>
        <VALUE>TARA_20090919Z</VALUE>
      </SAMPLE_ATTRIBUTE>
      <SAMPLE_ATTRIBUTE>
        <TAG>Sampling Station</TAG>
        <VALUE>TARA_005</VALUE>
      </SAMPLE_ATTRIBUTE>
      <SAMPLE_ATTRIBUTE>
        <TAG>Sampling Platform</TAG>
        <VALUE>SV Tara</VALUE>
      </SAMPLE_ATTRIBUTE>
      <SAMPLE_ATTRIBUTE>
        <TAG>Event Label</TAG>
        <VALUE>TARA_20090920T1045Z_005_EVENT_PUMP</VALUE>
      </SAMPLE_ATTRIBUTE>
    </SAMPLE_ATTRIBUTES>
  </SAMPLE>
</SAMPLE_SET>
```

Image 3.7.3: Example of ENA XML output file for a sample

### 3.7.1.7 Use of controlled vocabularies - which, where, how

Controlled Vocabularies are used and support building a search query as can be seen in the Query Builder component of: <https://www.ebi.ac.uk/ena/browser/advanced-search>

Also, it can be seen using the controlled Vocabularies endpoint by passing in a field name:

E.g. <https://www.ebi.ac.uk/ena/portal/api/controlledVocab?field=broker>

### **3.7.1.8 Data access policy - if yes, which and how deployed**

Terms of use for web and RESTful services can be found at:

<https://www.ebi.ac.uk/about/terms-of-use/>

GDPR notice for ENA Browser:

<https://www.ebi.ac.uk/data-protection/privacy-notice/ena-presentation>

GDPR notice for both RESTful services:

<https://www.ebi.ac.uk/data-protection/privacy-notice/embl-ebi-public-website>

### **3.7.1.9 Any web services and API's - URLs, function, how to operate**

Relevant API's are:

- ENA Data Discovery: <https://www.ebi.ac.uk/ena/portal/api/>
- ENA Data Retrieval: <https://www.ebi.ac.uk/ena/browser/api/>

How to use the API's and build machine-to-machine services can be found in the documentation of the ENA Portal API: <https://www.ebi.ac.uk/ena/portal/api/doc>

### **3.7.1.10 Any suggestions for aggregating data sets as collections in order to scale down number of too many entries and serve users with distinctive metadata entries as part of the planned Blue-Cloud data discovery and access service**

There are several options. Several of the ENA data classes serve to provide grouping of subordinate objects. Here, attributes of these grouping objects typically suffice to bring together what would otherwise be unwieldy numbers of records. Cases include:

- Assembly contig\_sets provide a summary view of an entire set of sequence contig records. E.g. <https://www.ebi.ac.uk/ena/browser/view/AOOT01000000> is a set of over a million contigs from a fish species.
- Study records services to connect large numbers of raw sequence data records with records of the samples that have been sequenced and the libraries that have been used for this sequencing. E.g. <https://www.ebi.ac.uk/ena/browser/view/PRJEB1787> represents a study in the Tara Oceans initiative that brings together 249 bacterial metagenomics sequence data sets.
- "Umbrella" study records bring together, with common attributes, a set of component studies (such as the above). E.g. <https://www.ebi.ac.uk/ena/browser/view/PRJEB402> brings together all studies under the Tara Oceans initiative.

### **3.7.1.11 Hosting environment**

ENA Portal API and ENA Browser API are Java based services deployed in Tomcat servers, running on EMBL-EBI managed data-centers, with redundancy, load balanced at the EMBL-EBI entry point.



### 3.7.1.12 Organisational aspects (main operator(s); data providers)

EMBL-EBI is managing ENA. As a provider of data services over the last four decades, ENA works with data providers from the smallest research groups to the world's largest sequencing facilities and a vast userbase of data consumers around the world.

### 3.7.1.13 Contact details

User support: <https://www.ebi.ac.uk/ena/browser/support>

Contacts for Blue-Cloud project:

- Suran Jayathilaka ([suranj@ebi.ac.uk](mailto:suranj@ebi.ac.uk))
- Guy Cochrane ([cochrane@ebi.ac.uk](mailto:cochrane@ebi.ac.uk))

### 3.7.1.14 Conclusion ELIXIR - ENA

EMBL-EBI operates API's for ENA discovery and ENA data retrieval which seem very suitable endpoints for connecting to the Blue-Cloud data discovery and access service. The ENA system contains many data types / classes and a huge volume of data, which are only partly marine related. Blue-Cloud should focus on data and information relevant for the marine domain and on data types such as samples and their analyses. A priority list needs to be determined as a next step. Moreover, the ENA system offers several algorithms / pipelines for processing data, which might be used in a 'smart' way for the Blue-Cloud. This also need to be analysed.

## 3.8 EuroBioImaging

EuroBioImaging is an EU ESFRI Research Infrastructure for Imaging Technologies in Biological and Biomedical Sciences. It provides access to a broad range of state-of-the-art technologies in biological and biomedical imaging for life scientists. In addition, it offers image data support and training for users and providers. It consists of 29 geographically distributed Node Candidates (specialised imaging facilities) and 36 imaging technologies are offered.

The BioImage Archive, as managed at EMBL-EBI, stores and distributes biological images that are useful to life-science researchers. It provides data archiving services to the broader bioimaging database community. The Archive was launched in July 2019, and currently is a collection of resources:

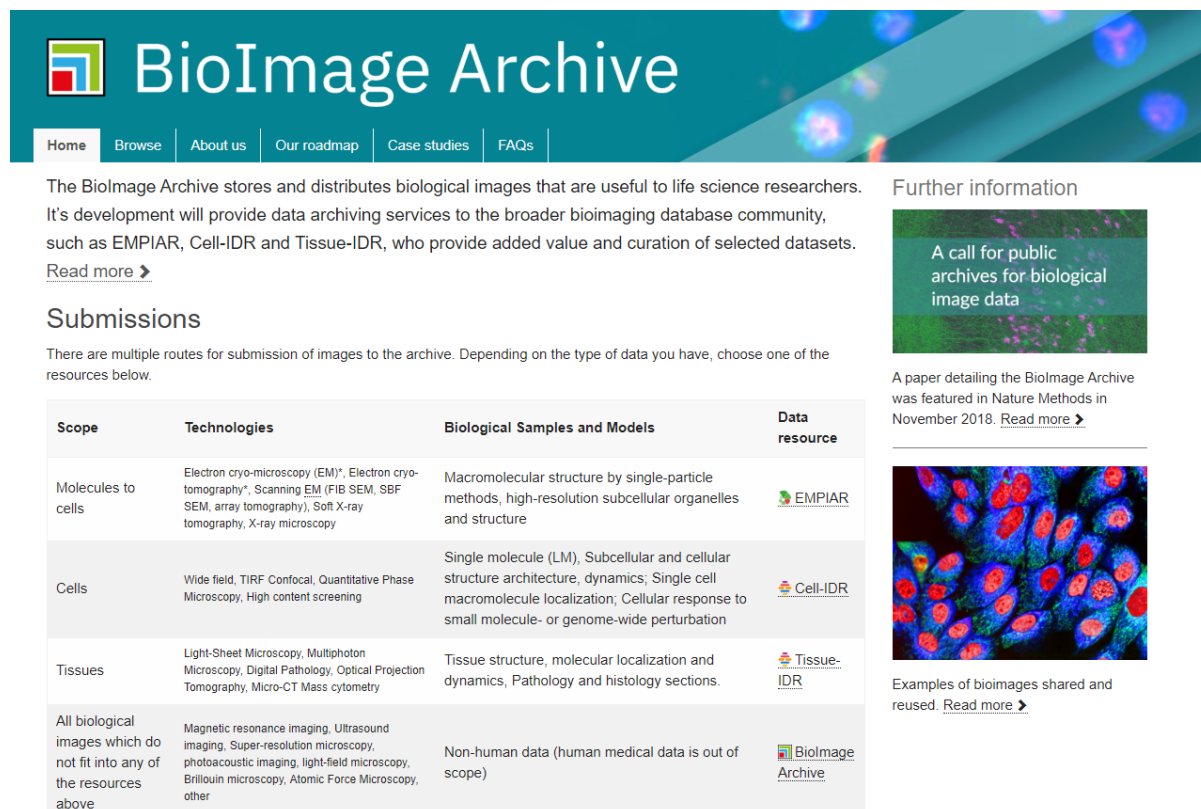
- EMPIAR for electron microscopy data,
- Cell-IDR and Tissue-IDR (Image Data Resource) for high quality, curated light microscopy data on cellular and tissue level respectively,
- BioStudies database that is a generic resource able to accept any other imaging datasets that do not fit into EMPIAR, Cell-IDR, or Tissue-IDR.

With content including marine microscopy data, the resource integrates imaging data with molecular and phenotype data from, e.g., ELIXIR data resources. It includes information on experimental protocols and outputs: parameters, analyses and variations observed in cells and

features under different experimental variables/parameters. All data is available at the portal. All software for building and running the archive and reading metadata is open source and available on GitHub.

### 3.8.1 Data discovery and access service component

The BioImage Archive has discovery for each of its database modules EMPIAR, Cell-IDR and Tissue-IDR, and BioStudies.



The BioImage Archive stores and distributes biological images that are useful to life science researchers. It's development will provide data archiving services to the broader bioimaging database community, such as EMPIAR, Cell-IDR and Tissue-IDR, who provide added value and curation of selected datasets. [Read more](#)

### Submissions

There are multiple routes for submission of images to the archive. Depending on the type of data you have, choose one of the resources below.

Scope	Technologies	Biological Samples and Models	Data resource
Molecules to cells	Electron cryo-microscopy (EM)*, Electron cryo-tomography*, Scanning EM (FIB SEM, SBF SEM, array tomography), Soft X-ray tomography, X-ray microscopy	Macromolecular structure by single-particle methods, high-resolution subcellular organelles and structure	<a href="#">EMPIAR</a>
Cells	Wide field, TIRF Confocal, Quantitative Phase Microscopy, High content screening	Single molecule (LM), Subcellular and cellular structure architecture, dynamics; Single cell macromolecule localization; Cellular response to small molecule- or genome-wide perturbation	<a href="#">Cell-IDR</a>
Tissues	Light-Sheet Microscopy, Multiphoton Microscopy, Digital Pathology, Optical Projection Tomography, Micro-CT Mass cytometry	Tissue structure, molecular localization and dynamics, Pathology and histology sections.	<a href="#">Tissue-IDR</a>
All biological images which do not fit into any of the resources above	Magnetic resonance imaging, Ultrasound imaging, Super-resolution microscopy, photoacoustic imaging, light-field microscopy, Brillouin microscopy, Atomic Force Microscopy, other	Non-human data (human medical data is out of scope)	<a href="#">BioImage Archive</a>

**Further information**

A call for public archives for biological image data

A paper detailing the BioImage Archive was featured in Nature Methods in November 2018. [Read more](#)

Examples of bioimages shared and reused. [Read more](#)

*Image 3.8.1: Homepage of BioImage Archive with access to EMPIAR, Cell-IDR, Tissue-IDR, and BioStudies*

#### 3.8.1.1 Name

BioImage Archive

#### 3.8.1.2 Web address

Index of BioImage Archive datasets in EMPIAR and BioStudies (currently does not include Cell-IDR or Tissue-IDR), and access to datasets in BioStudies:

<https://www.ebi.ac.uk/biostudies/BioImages/studies>

Discover and access data in EMPIAR:

<https://www.ebi.ac.uk/pdbe/emdb/empiar/>

Discover and access data in Cell-IDR and Tissue-IDR:

<https://idr.openmicroscopy.org/cell/>

<https://idr.openmicroscopy.org/tissue/>

### 3.8.1.3 Types and number of data sets and/or data products

The BioImage Archive contains the following number of data sets as of November 2019:

- EMPIAR: 265 datasets
- BioStudies: 434 datasets
- Cell-IDR: 57 datasets
- Tissue-IDR: 17 datasets

See <https://idr.openmicroscopy.org/about/studies.html> for IDR data summary.

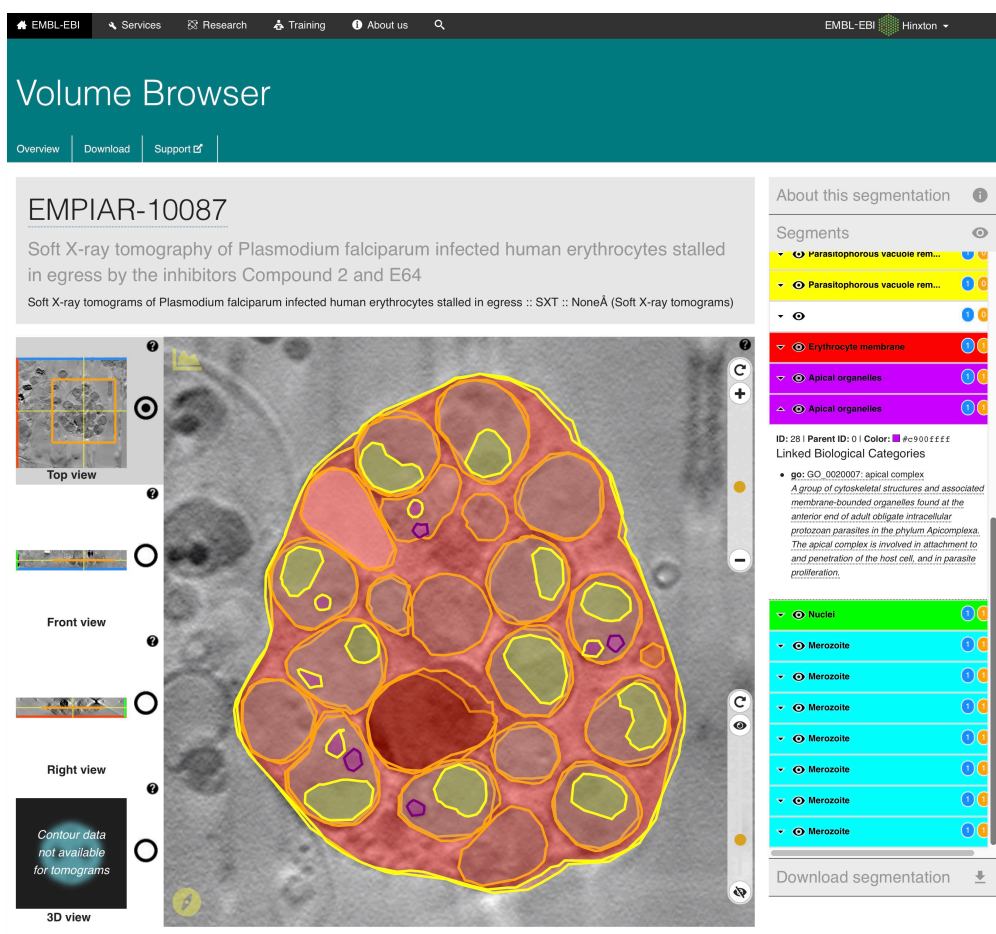


Image 3.8.2: Example of image at EMPIAR

### 3.8.1.4 Discovery and access mechanisms - how does it function

EMPIAR has web UI and a dataset detail access API. Documentation can be found at:

<https://www.ebi.ac.uk/pdbe/emdb/empiar/api/documentation/>

It supports FTP, Aspera, and Globus downloads.

IDR has web UI and a JSON-based API for accessing all datasets, thumbnails and metadata.

Documentation can be found at:

<https://idr.openmicroscopy.org/about/api.html>

It supports Aspera downloads.

### **3.8.1.5 Metadata format(s) - short overview and references to detailed documentation**

EMPIAR has its own XML metadata format.

IDR has metadata access via API (standard OMERO.web API and newly developed plugins like OMERO.mapr for finding images via attributes linked to images).

BioStudies has a very generic, custom metadata format, which can be expressed as a tab-delimited file, JSON, or XML. Documentation for the tab-delimited version (from the submission viewpoint) is available from <https://www.ebi.ac.uk/biostudies/submit>.

### **3.8.1.6 Data format(s) - short overview and references to detailed documentation**

EMPIAR provides image data in the formats in which they are uploaded; it is recommended to use common formats in the field including MRC, MRCS, TIFF, DM4, IMAGIC, SPIDER, MRC FEI and RAW FEI.

IDR accepts and allows downloading data in formats readable by Bio-Formats library.

BioStudies accepts and distributes data in any format; it is recommended to use imaging formats readable by the Bio-Formats library.

### **3.8.1.7 Use of controlled vocabularies - which, where, how**

Community discussions about imaging metadata, including controlled vocabularies, have started, but the current implementations of both data capture and distribution are not relying on these. One exception is Cellular Microscopy Phenotype Ontology (<https://www.ebi.ac.uk/cmpo/>) and Experimental Factor Ontology (<https://www.ebi.ac.uk/efo/>) used in IDR.

### **3.8.1.8 Data access policy - if yes, which and how deployed**

EMPIAR: CC0 license, terms of use: <https://www.ebi.ac.uk/about/terms-of-use/>

IDR: different CC licences used, on per-dataset basis: CC0, CC-BY 4.0, CC-BY-SA 3.0, CC-BY-NC 4.0, CC-BY-NC-ND 4.0, CC BY-NC-SA 3.0 .

BioStudies: terms of use: <https://www.ebi.ac.uk/about/terms-of-use/>; datasets imported from Journal of Cell Biology (<https://www.ebi.ac.uk/biostudies/BioImages/studies?facet.project=jcb>) are under a range of CC licences.

### **3.8.1.9 Any web services and API's - URLs, function, how to operate**

EMPIAR - REST API: <https://www.ebi.ac.uk/pdbe/emdb/empiar/api/documentation/>

IDR: <https://idr.openmicroscopy.org/about/api.html>

BioStudies - REST API: <https://www.ebi.ac.uk/biostudies/help>

#### **3.8.1.10 Any suggestions for aggregating data sets as collections in order to scale down number of too many entries and serve users with distinctive metadata entries as part of the planned Blue-Cloud data discovery and access service**

As in ENA, “umbrella datasets” can be easily created in BioStudies, bringing together several imaging datasets with other data types.

#### **3.8.1.11 Hosting environment**

BioStudies: UI and APIs - Java based services deployed in Tomcat servers, running on EMBL-EBI managed data-centers, with redundancy, load balanced at the EMBL-EBI entry point

IDR: see <https://idr.openmicroscopy.org/about/deployment.html>

#### **3.8.1.12 Organisational aspects (main operator(s); data providers)**

EMPIAR: developed and run by the Cellular Structure and 3D Bioimaging team, EMBL-EBI; community data depositions.

IDR: developed and run by the OME Consortium at the University of Dundee (PI Jason Swedlow), with EMBL-EBI participation; hosting on EMBL-EBI Embassy cloud; community data depositions.

BioStudies: developed and run by the Functional Genomics Software Development team, EMBL-EBI; community data depositions.

#### **3.8.1.13 Contact details**

User support and contact under Blue-Cloud:

- Stéphane Pesant <[spesant@marum.de](mailto:spesant@marum.de)>
- Guy Cochrane ([cochrane@ebi.ac.uk](mailto:cochrane@ebi.ac.uk))

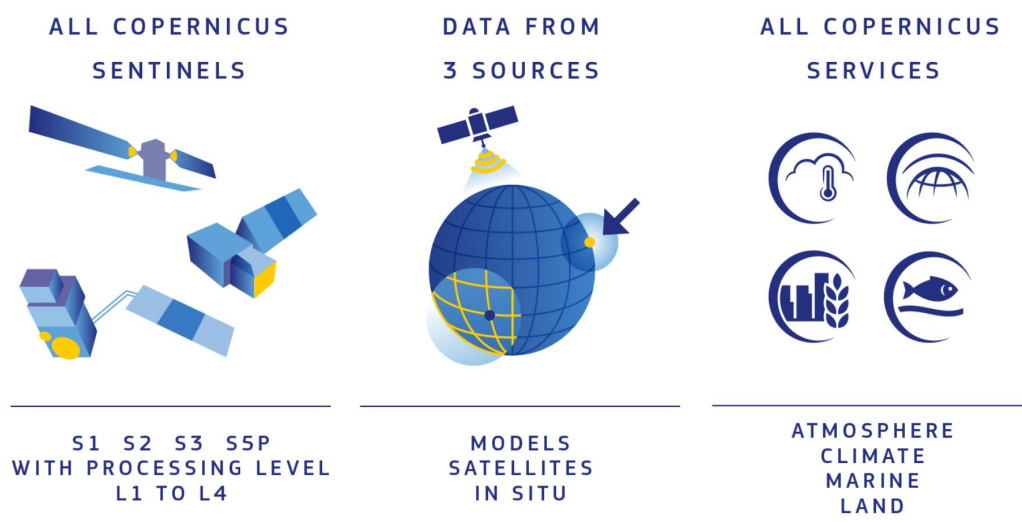
#### **3.8.1.14 Conclusions EuroBioImaging**

As part of EuroBioImaging the BioImage Archive is operated. This consists of 4 separate databases (EMPIAR; Cell-IDR; Tissue-IDR; BioStudies) with different metadata and data models, and different search and access API's. Content is only partly marine related. Blue-Cloud should focus on images and databases relevant for the marine domain. This should be analysed as a next step in order to determine if all databases need to be coupled to the Blue-Cloud.

## **3.9 WEKEO**

WEKEO is one of the 5 Copernicus DIAS (Data and Information Access Services). The overarching objective of DIAS is to enhance access to Copernicus data and information for further use in an efficient computing environment implementing the paradigm of “bringing the user to the data”, as one condition for unlocking the potential value of Copernicus for innovation, science, new business, implementation of public policies and economic growth.





*Image 3.9.1: Overview of planned offering of WEkEO portal*

Considering the well-structured user communities and the strong relationship existing between EUMETSAT (operator and data provider for Jason-3, Sentinel-3, Sentinel-4, Sentinel-5, Jason-CS/Sentinel-6 satellites), Mercator-Océan (Copernicus Marine Environment Monitoring Service (CMEMS)) and ECMWF (Copernicus Atmosphere Monitoring Service (CAMS), Copernicus Climate Change Service (C3S)), the 3 organisations have joined their efforts to implement one instance of a DIAS in partnership, namely WEkEO. WEkEO is the service for marine environmental data, virtual environments for data processing, and skilled user support. WEkEO will have a public and free part for discovery and access to data and data products, while it will also have a commercial part with various analysis applications and cloud space. The WEkEO commercial exploitation has been awarded to the consortium led by THALES ALENIA SPACE and CloudFERRO.

### **3.9.1 Data discovery and access service component**

The WEkEO portal with its discovery and access services is currently under development and will be launched in the first quarter of 2020.

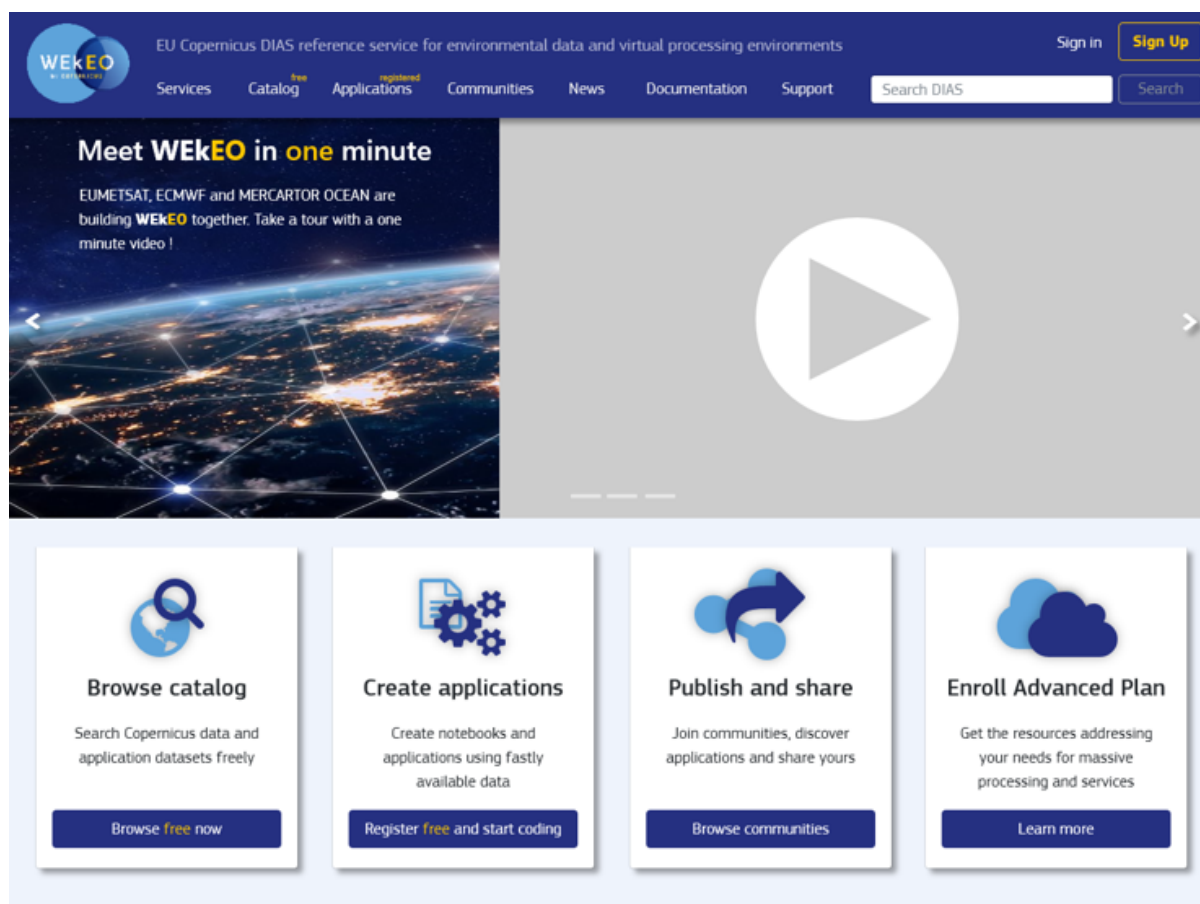


Image 3.9.2: Preview of the WEkEO portal (to be released in Q1 2020).

### 3.9.1.1 Name

Harmonized Data Access (HAD)

### 3.9.1.2 Web address

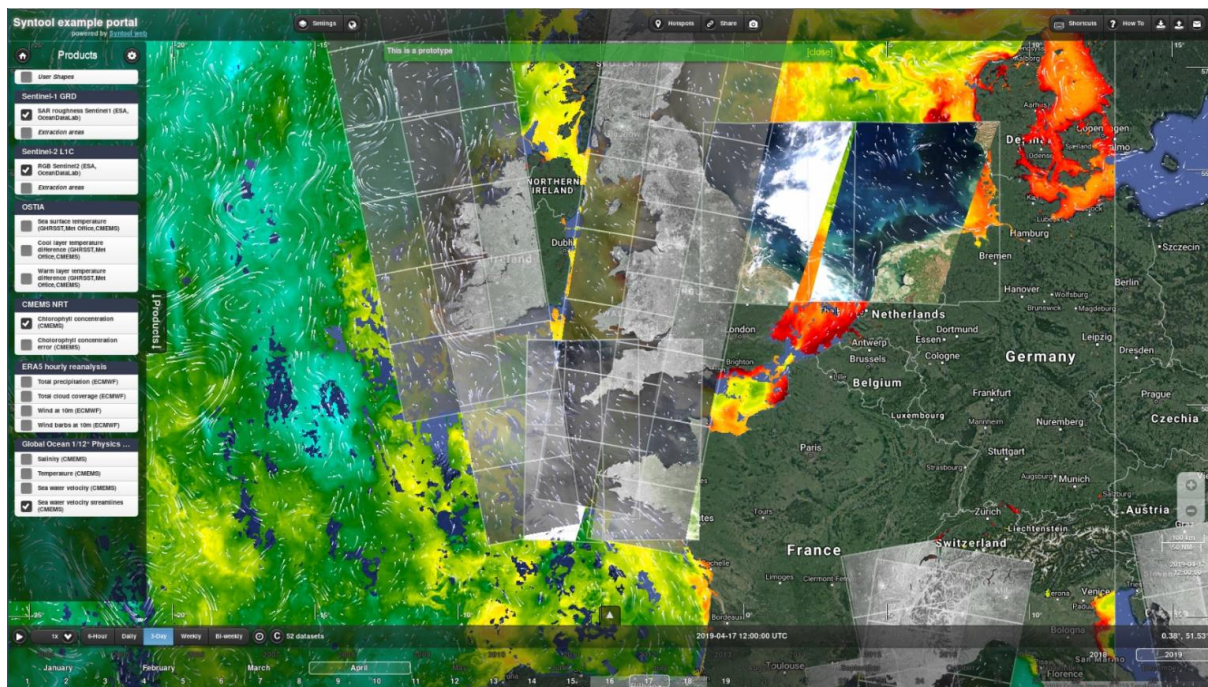
Most probably: [https://www.wekeo.eu/harmonised\\_data\\_access\\_API](https://www.wekeo.eu/harmonised_data_access_API)

Not yet active. Planned for Q1 2020.

### 3.9.1.3 Types and number of data sets and/or data products

The HAD will provide discovery and access to:

- All the Sentinel satellite data sets: S1, S2, S3 Marine, S3 Land, S5P
- Main Copernicus Service Data products from:
  - Copernicus Marine Service: CMEMS
  - Copernicus Atmospheric Service: CAMS
  - Copernicus Climate Service: C3S
  - Copernicus Land Service: CLMS



*Image 3.9.3: Example of data aggregation*

### 3.9.1.4 Discovery and access mechanisms - how does it function

The HAD will have a single access protocol (REST API) whereby scaling and evolution of codes will be made easy. It will provide easy & fast access to subsetting attributes in hundreds of datasets from difference Copernicus sources.

Information is provided in Swagger : More details are not yet available. Planned for Q1 2020.

<http://wekeo-broker.eumetsat-dpi.wekeo-dev.cloudferro.com:8080/databroker/ui/#/>

More details will follow in Q1 2020.

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

Details not yet available. Planned for Q1 2020.

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

Details not yet available. Planned for Q1 2020.

### 3.9.1.7 Use of controlled vocabularies - which, where, how

Details not yet available. Planned for Q1 2020.

### 3.9.1.8 Data access policy - if yes, which and how deployed

Copernicus data policy will apply: Free and Open access to all data and data products.

### **3.9.1.9 Any web services and API's - URLs, function, how to operate**

The Harmonised Data Access (HDA) API will allow uniform access to the whole WEkEO catalogue, including subsetting and downloading functionalities. The HDA API will be REST-based. More details are not yet available. Planned for Q1 2020.

### **3.9.1.10 Any suggestions for aggregating data sets as collections in order to scale down number of too many entries and serve users with distinctive metadata entries as part of the planned Blue-Cloud data discovery and access service**

Details not yet available. Planned for Q1 2020.

### **3.9.1.11 Hosting environment**

WEkEO will be hosted on a distributed infrastructure, centred on the CloudFERRO cloud in Warsaw, with ramification in EUMETSAT; Mercator and ECMWF.

### **3.9.1.12 Organisational aspects (main operator(s); data providers)**

WEkEO will be managed by its partners: EUMETSAT, ECMWF and Mercator Ocean International. WEkEO will be operated by Thales Alenia Space and CloudFERRO.

Sentinel Data will be provided from CloudFERRO local copy and EUMETSAT for S3 Marine, while the Copernicus services data products will come from their original data distributors via dedicated fiber.

### **3.9.1.13 Contact details**

For Blue-Cloud:

- Alain Arnaud (MOI) (alain.arnaud@mercator-ocean.fr)
- Renaud Dussurget (MOI) (renaud.dussurget@mercator-ocean.fr)

### **3.9.1.14 Conclusion WEkEO**

WEkEO is under development and will feature the Harmonised Data Access (HDA) API which will allow uniform access to the whole WEkEO catalogue of Sentinel satellite images and Copernicus data products from CMEMS, C3S, CAMS, and CLMS, including subsetting and downloading functionalities. The HDA API will be REST-based. The WEkEO discovery and access services are planned for release in Q1 2020. As soon as launched, further details for WEkEO need to be gathered and included in the next deliverable D2.2.

## **3.10 ICOS – Marine**

ICOS is an international organisation of thirteen European member countries and over 130 greenhouse gas measurement stations aimed at quantifying and understanding the greenhouse gas balance of Europe and neighbouring regions. ICOS data is made available at the Carbon Portal, a one-stop shop for all ICOS data products. The Ocean Thematic Centre is one of four central facilities



within the European research infrastructure Integrated Carbon Observation System (ICOS). The marine element of ICOS provides long-term oceanic observations, which are required to understand the present state and better predict future behaviour of the global carbon cycle and climate-relevant gas emissions. The Ocean Thematic Centre currently coordinates twenty-one ocean stations from seven countries monitoring carbon uptake and fluxes in the North Atlantic, Nordic Seas, Baltic, and the Mediterranean Sea. Measuring methods include sampling from research vessels, moorings, buoys, and commercial vessels that have been equipped with state-of-the-art carbonate system sensors. The objective is to ensure high quality measurements of greenhouse gas concentrations that are independent, transparent and reliable. In turn, this monitoring system will support governments in their efforts to mitigate climate change as well as holding them accountable for reaching their mitigation targets.

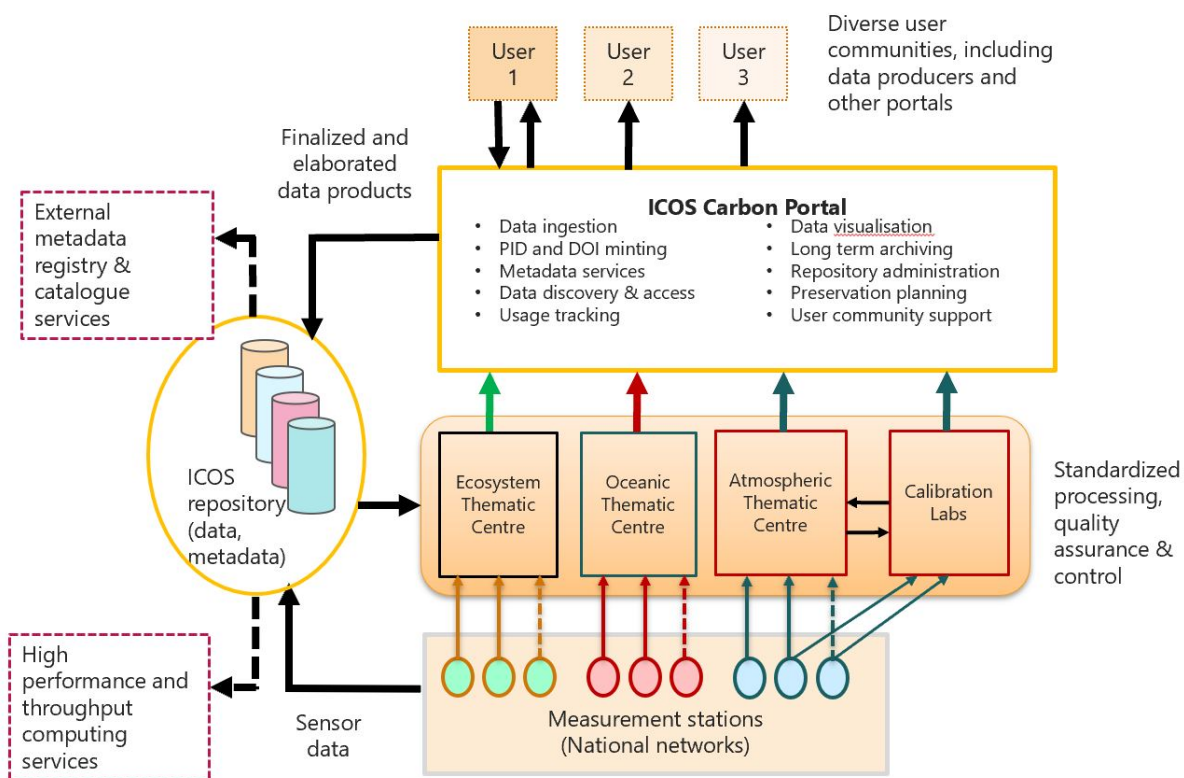


Image 3.10.1: Overview of ICOS data flow and organisation

### 3.10.1 Data discovery and access service component

Data from official certified ICOS stations can be accessed at the ICOS Carbon Portal and on an international scale including non ICOS data at the SOCAT Portal.

#### 3.10.1.1 Name

ICOS Carbon Portal  
SOCAT Portal



### 3.10.1.2 Web address

<https://www.icos-cp.eu/data>  
<https://data.icos-cp.eu/portal/>  
<https://meta.icos-cp.eu/sparqlclient/>  
<https://www.socat.info>

### 3.10.1.3 Types and number of data sets and/or data products

The ICOS Carbon Portal provides observation data from over 130 greenhouse gas measurement stations. Through the SOCAT portal there are around 6000 trajectories available from 1957-2018 with 26 million data values.

### 3.10.1.4 Discovery and access mechanisms - how does it function

The ICOS Carbon Portal facilitates an UI and SPARQL endpoint for searching in metadata. Not all datasets presented in the Data Portal are directly available. Raw data that was not quality assessed and checked is only available on request both for preview and download, but all metadata information is present. Use is made of a shopping cart system for downloading data sets. This requires accepting the ICOS Data Policy and Data Licensing Agreements first and registering an account.

SOCAT data can be accessed via various mechanisms:

- File by file via PANGAEA
- ZIP folder via NOAA NCEI
- At SOCAT portal via ERDDAP, as ODV and as ZIP files
- SOCAT Data is also distributed via CMEMS and ICOS Carbon Portal.

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

The metadata of the ICOS Carbon Portal are based on ontology, whereby all elements have (linked) URIs. Documented in RDF at: <http://meta.icos-cp.eu/ontologies/cpmeta/>

SOCAT uses Enhanced metadata reporting for carbon data (fCO<sub>2</sub>): details can be found in the respective Best Practice 'Indicator Methodology for SDG 14.3.1: Indicator Description 14.3.1 – Average marine acidity (pH) measured at agreed suite of representative sampling stations.' See: <https://www.oceanbestpractices.net/handle/11329/1132>

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

ODV, NetCDF, and various other file formats

### 3.10.1.7 Use of controlled vocabularies - which, where, how

ICOS Carbon Portal metadata are documented in RDF at: <http://meta.icos-cp.eu/ontologies/cpmeta/>  
Standardised vocabs (BODC/NERC, CF) will be adopted from 2020 to be compliant with other European data.

SOCAT uses CF in the NetCDF files.

#### **3.10.1.8 Data access policy - if yes, which and how deployed**

The data license of the Carbon Portal is described at: <https://data.icos-cp.eu/licence>

SOCAT uses CCBY 4.0

#### **3.10.1.9 Any web services and API's - URLs, function, how to operate**

ICOS Carbon Portal: there is a SPARQL endpoint and RESTful API for metadata. For data access there is a DAP service.

SOCAT:

[https://ferret.pmel.noaa.gov/socat/erddap/tabledap/socat\\_v2019\\_fulldata.html](https://ferret.pmel.noaa.gov/socat/erddap/tabledap/socat_v2019_fulldata.html)

#### **3.10.1.10 Any suggestions for aggregating data sets as collections in order to scale down number of too many entries and serve users with distinctive metadata entries as part of the planned Blue-Cloud data discovery and access service**

No suggestions. Will need to analysed further.

#### **3.10.1.11 Hosting environment**

ICOS Carbon Portal is supported by EUDAT for long-term archiving (B2SAFE).

SOCAT data is archived by Bjerknes Climate Data Centre which hosts the data management system for ICOS OTC at the University of Bergen.

#### **3.10.1.12 Organisational aspects (main operator(s); data providers)**

ICOS Carbon Portal is managed by University of Lund.

SOCAT data is managed by the Bjerknes Climate Data Centre.

#### **3.10.1.13 Contact details**

For Blue-Cloud purpose:

- SOCAT: Benjamin Pfeil (University of Bergen) ([Benjamin.Pfeil@uib.no](mailto:Benjamin.Pfeil@uib.no))
- ICOS Carbon Portal: Alex Vermeulen (University of Lund) ([alex.vermeulen@icos-ri.eu](mailto:alex.vermeulen@icos-ri.eu))

#### **3.10.1.14 Conclusion ICOS-Marine**

This concerns two relevant portals: ICOS Carbon Portal with data discovery and access and SOCAT portal with data products. Several services are available for both. Some more detail is needed about metadata and data formats, in particular because the ICOS Carbon Portal is upgrading its metadata format and adopting vocabularies as part of the ENVRI-FAIR project. Web services for discovery are existing, but it might be needed to specify and develop API's for data access as part of the Blue-Cloud activities.

## 4 Conclusions and planned follow-up

The pilot Blue-Cloud project aims at federating initially in total 10 blue data infrastructures. Each of these existing infrastructures have been described in this deliverable D2.1, in particular with a focus on their current data discovery and access mechanisms, if existing. From the descriptions and discussions, among others at the first TCom meeting in January 2020, it appears that a number of blue data infrastructures do not have to be federated with direct interfacing to the Blue-Cloud data discovery and access service, but indirectly as some are or will be coupled to another of the blue data infrastructures and that way it will be arranged that their contents will also be present in the Blue-Cloud data discovery and access service. In addition, activities are needed in several cases for providing API's which are suited for serving the planned Blue-Cloud data discovery and access service. The following table gives the conclusions of this initial analysis per blue data infrastructure.

Blue Data Infrastructure	Coupling to Blue-Cloud	Conclusions
SeaDataNet	Direct	SeaDataNet operates the CDI data discovery and access service. For exchange to Blue-Cloud this already features an INSPIRE compliant API at aggregate metadata level. Still to be specified and developed is a data access API with Marine-ID authentication, capable of processing data requests at aggregate metadata level.
EMODnet Bathymetry	Indirect for data and Direct for products	EMODnet Bathymetry makes use of the SeaDataNet CDI data discovery and access service. Therefore, no separate solutions need to be developed for Blue-Cloud. The highly popular EMODnet Digital Terrain Model (DTM) data product is relevant for Blue-Cloud purposes and can be used through existing OGC web services.
EMODnet Chemistry	Indirect for data and Direct for products	EMODnet Chemistry makes use of the SeaDataNet CDI data discovery and access service. Therefore, no separate solutions need to be developed for Blue-Cloud. The aggregated, harmonized and validated data collections for eutrophication, contamination, acidification and marine litter, as regularly produced by EMODnet Chemistry, are also relevant for Blue-Cloud purposes. Developments are underway for establishing an API and GUI for facilitating sub-setting and retrieval of these data collections. Once operational, this service will provide an additional channel to be added to the Blue-Cloud data discovery and access service.
EuroArgo - Argo	Direct	EuroArgo operates a number of web services for discovery and access to the ArgoFloat data sets. These can be used for the first release of the Blue-Cloud data discovery and access service. EuroArgo is developing advanced services as part of the ENVRI-FAIR, EIOSC-HUB, and EA-RISE

Blue Data Infrastructure	Coupling to Blue-Cloud	Conclusions
		projects, which should be followed closely as near-future candidate for coupling to the Blue-Cloud.
EurOBIS – EMODnet Biology	Direct	EurOBIS – EMODnet Biology operates a number of web services for discovery and access to the EurOBIS data sets. Of these, the endpoint of the Integrated Publishing Toolkit (IPT) seems to be most suited for connecting to the Blue-Cloud data discovery and access service.
EcoTaxa	Indirect	As part of the Blue-Cloud, EcoTaxa metadata – data will be integrated in EurOBIS – EMODnet Biology by API which is under development. The coupling to the Blue-Cloud data discovery and access service will then be provided through EurOBIS – EMODnet Biology.
ELIXIR – ENA	Direct	EMBL-EBI operates API's for ENA discovery and ENA data retrieval which seem very suitable endpoints for connecting to the Blue-Cloud data discovery and access service. The ENA system contains many data types / classes and a huge volume of data, which are only partly marine related. Blue-Cloud should focus on data and information relevant for the marine domain and on data types such as samples and their analyses. A priority list needs to be determined as a next step. Moreover, the ENA system offers several algorithms / pipelines for processing data, which might be used in a 'smart' way for the Blue-Cloud. This also needs to be analysed.
EuroBioImaging	Direct	As part of EuroBioImaging the BioImage Archive is operated. This consists of 4 separate databases (EMPIAR; Cell-IDR; Tissue-IDR; BioStudies) with different metadata and data models, and different search and access API's. Content is only partly marine related. Blue-Cloud should focus on images and databases relevant for the marine domain. This should be analysed as a next step in order to determine if all databases need to be coupled to the Blue-Cloud.
WEkEO	Direct	WEkEO is under development and will feature the Harmonised Data Access (HDA) API which will allow uniform access to the whole WEkEO catalogue of Sentinel satellite images and Copernicus data products from CMEMS, C3S, CAMS, and CLMS, including subsetting and downloading functionalities. The HDA API will be REST-based. The WEkEO discovery and access services are planned for release in Q1 2020. As soon as launched, further details for WEkEO need to be gathered and included in the next deliverable D2.2.

Blue Data Infrastructure	Coupling to Blue-Cloud	Conclusions
ICOS-Marine	Direct	This concerns two relevant portals: ICOS Carbon Portal with data discovery and access and SOCAT portal with data products. Several services are available for both. Some more detail is needed about metadata and data formats, in particular because the ICOS Carbon Portal is upgrading its metadata format and adopting vocabularies as part of the ENVRI-FAIR project. Web services for discovery are existing, but it might be needed to specify and develop API's for data access as part of the Blue-Cloud activities.

This initial description and evaluation will be followed by a deeper analysis, detailing the technical specifications of the Blue-Cloud data discovery and access service, and possible developments required at and by each of the blue data infrastructures. This analysis, technical specification and workplan for the implementation and deployment of the Blue-Cloud data discovery and access service will be documented in Deliverable D2.2 at M8. For this analysis, there will be cooperation and synergy with the activities and results of the ENVRI-FAIR project as several Blue Cloud marine data infrastructures (Euro-Argo, ICOS-Marine, and SeaDataNet) are involved in both projects and as ENVRI-FAIR will analyse and improve the FAIRness of their data management services. CNR-IIA, MARIS, and University of Amsterdam will interact with the Data & Access service providers for mapping their metadata to the common metadata model, establishing a suitable aggregation of data collections, and arranging metadata harvesting services.

D2.2 will be followed by further development and implementation activities at central level and at each of the blue data infrastructures. For instance, CNR-IIA will configure and deploy their GEODAB service to function as Blue-Cloud metadata brokerage service, installing adaptors for each of the Data & Access services following the earlier mappings, and then initialising dynamic harvesting of the metadata for data collections. The resulting Blue-Cloud metadata catalogue will be dynamic and operationally published by CNR-IIA as CSW, OAI-PMH and SPARQL services, on top of which MARIS will develop and deploy the Blue-Cloud catalogue GUI.

As a next step, MARIS will develop an operational Blue-Cloud data brokerage, that will interact with the partly existing and partly to be developed API's of the blue data infrastructures. In dialogue with MARIS each Data & Access provider (see Task 2.1) will specify, develop and deploy their API by which data collections in sync with their agreed metadata entries can be retrieved by the data brokerage service, taking into account the local data access mechanisms. One part is a shopping mechanism, directly linked to the Blue Cloud metadata catalogue (resulting from Task 2.1), and individual adaptors dealing with the API's of Data & Access providers. The shopping mechanism will include AAA services for keeping track of requests and facilitating the delivery services to customers. Another part is a shopping ledger for users and Data & Access providers to keep track of shopping requests and progress by shopping adaptors for fulfilling the requests. The shopping components will be derived from the



services that MARIS already has developed in the SeaDataCloud project as part of the SeaDataNet CDI service.

A third part is making arrangements for temporary storage of retrieved datasets and delivery to users (by downloading) as well as to the Blue-Cloud VRE. This will be developed by EUDAT.

Integration of the Blue-Cloud data catalogue service and data broker will establish the Blue-Cloud data discovery and access service with GUI and API which is planned for launch end M17. Overall, the steps from specification to development to integration to operational deployment to acceptance testing will be followed.