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

D2.1 – Existing DD&AS and Blue Data Infrastructures – Review and Specifications for Optimisation Report

Work Package	WP2 | FAIR compliant Discovery and Access services for marine domains & beyond
Task	T2.1 | Optimising the functioning of the Data Discovery & Access Service
Lead author	Dick M.A. Schaap (MARIS)
Contributors	Enrico Boldrini (CNR-IIA), Roberto Roncella (CNR-IIA), Peter Thijsse (MARIS), Paul Weerheim (MARIS), Serge vd Horst (MARIS), Robin Kooyman (MARIS), Bert Broeren (MARIS), Tjerk Krijger (MARIS), Patricia Cabrera (VLIZ), Bart Vanhoorne (VLIZ), Katrina Exter (VLIZ), Mark Portier (VLIZ), Alexandra Kokkinaki (NOC-BODC), Gwenaelle Moncoiffe (NOC-BODC), Dominique Obaton (IFREMER), Thierry Carval (IFREMER), Christine Coatanoan (IFREMER), Erwan Bodere (IFREMER), Alessandra Giorgetti (OGS), Eugenia Molina (OGS), Gwenael CAER (IFREMER), Sissy Iona (HCMR), Beatrice Chiavarini (CINECA), Chris Ariyo (CSC), Guy Cochrane (EML-EBI), Stéphane Pesant (EMBL-EBI), Rob Finn (EMBL-EBI), Suran Jayathilaka (EMBL-EBI), Lili Meszaros (EMBL-EBI), Jean-Olivier Irisson (SU), Béatrice Caraveo (SU), Steve Jones (UIB), Benjamin Pfeil (NORCE), Richard Sanders (NORCE), Alex Vermeulen (LU), Angeliki Adamaki (LU)
Peer reviewers	Pasquale Pagano (CNR-ISTI), Sara Pittonet Gaiarin (Trust-IT), Enrico Boldrini (CNR-IIA)
Version	V0.1
Due Date	31/10/2023
Submission Date	31/10/2023

Funded by the European Union

Dissemination Level

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

Version History

Revision	Date	Editors	Comments
0.1	24/10/2023	Dick M.A. Schaap (MARIS)	First draft
0.2	27/10/2023	Pasquale Pagano (CNR-ISTI), Enrico Boldrini (CNR-IIA)	Internal review
0.3	30/10/2023	Sara Pittonet Gaiarin (Trust-IT)	Internal review
0.4	31/10/2023	Pasquale Pagano (CNR-ISTI)	Final version and upload

Glossary of terms

Item	Description
API	Application Programming Interface
BDI	Blue Data Infrastructure
CDI	Common Data Index (SeaDataNet)
CELERY	A simple, flexible, and reliable distributed system to process vast amounts of messages
CMEMS	Copernicus Marine Environment Monitoring Service
CSW	Catalogue Service for the Web (OGC standard)
DAB	Discovery and Access Broker
DDAS	Data Discovery & Access Service
EcoTaxa	Web application dedicated to the visual exploration and the taxonomic annotation of images focused on planktonic biodiversity

Item	Description
EBWBL	EMODnet Bathymetry World Base Layer
ELIXIR-ENA	European Nucleotide Archive (ELIXIR service)
EML	Ecological Metadata Language (EurOBIS standard)
EMODnet	European Marine Observation and Data network
EOSC	European Open Science Cloud
ERDDAP	Environmental Research Division's Data Access Program
EurOBIS	European node of Ocean Biogeographic Information System
Euro-Argo	European contribution to Argo program
FLASK	A micro web framework written in Python
GBIF	Global Biodiversity Information Facility
GDAC	Global Data Assembly Centre
GUI	Graphical User Interface
HAD	Harmonised Data Access (WekEO service)
ICOS	Integrated Carbon Observing System
ISO	International Organization for Standardization
JSON	JavaScript Object Notation
MSFD	Marine Strategy Framework Directive
NGINX	Load Balancer, Reverse Proxy
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OBIS	Ocean Biogeographic Information System
OGC	Open Geospatial Consortium
OPENSTACK	Cloud Computing Software Platform
Redis	Open source in-memory data structure store
SeaDataNet	pan-European network for marine and ocean data management
SOCAT	Surface Ocean CO2 Atlas
SPARQL	SPARQL Protocol and RDF Query Language
SQL	Structured Query Language
TSC	Technical and Scientific Committee
TLS	Transport Layer Security
UI	User Interface
VLab	Virtual Laboratory
VRE	Virtual Research Environment

Funded by the European Union
Blue-Cloud 2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227

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

Keywords

EOSC; Blue-Cloud ; Data Discovery ; Data Access ; Federated Discovery ; Federated Access ; Subsetting

Disclaimer

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

EXECUTIVE SUMMARY

The **Blue Cloud Data Discovery and Access Service** (**DD&AS**) is one of the components of the Blue-Cloud technical framework. It serves federated discovery and access to Blue Data Infrastructures (BDIs) for external users and also interacts with the Blue-Cloud Virtual Research Environment (the component federating computing platforms and analytical services) for populating the VRE data pool. Within the framework of the H2020 pilot Blue-Cloud project, the first operational release of the Blue-Cloud DD&AS was deployed. This service is federating in total 8 BDIs and with 9 data services.

The implementation of the Blue Cloud DD&AS largely depends on machine-to-machine interactions between central components of the DD&AS and web services / APIs as managed and operated by the BDIs. In practice, the development has been quite challenging as each of the BDIs is offering web services for discovery and access which are different in functionality and not always straightforward. In the Blue-Cloud 2026 project it is strived for further optimisation of the DD&AS, in particular, where possible, by improving the FAIRness of the services and contents as provided by the federated BDIs. In addition, it is explored how the overall functionality of the DD&AS might be expanded.

In this Deliverable, first of all, an overview is given of the current architecture and functionality of the Blue-Cloud DD&AS to provide good insight in the DD&AS conceptual principles. Then per BDI an overview is given of their current federation, followed by an assessment of possible improvements, which should be undertaken by and/or with each BDI. These analyses and assessment have been made via a continuous dialogue between MARIS and CNR-IIA as core DD&AS developers with each of the BDIs and associated technical providers. This started at the Project Kick-Off meeting, 13–15 February 2023 in Pisa – Italy, during a dedicated WP2 session, and the first Technical and Scientific Committee (TSC) meeting, organised on 28-29 March 2023 in Amsterdam – Netherlands. At these meetings the common principles for optimisation of the DD&AS were discussed and agreed. Since then, the core developers have been reviewing the current deployments, supported by email exchanges with BDIs, where needed. This has been followed by bilateral web conferences with managers of each of the BDIs between August and October 2023 to discuss findings and ideas. Finally, this has resulted in a series of proposed improvements for each of the BDIs and the central DD&AS core service. These will be worked out and taken into development in the coming period as part of WP2 activities.

TABLE OF CONTENTS

Funded by the European Union
Blue-Cloud 2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227

Funded by the European Union
Blue-Cloud 2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227

TABLE OF FIGURES

Funded by the European Union

Blue-Cloud 2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227

TABLE OF TABLES

1. Introduction

The **Blue Cloud Data Discovery and Access Service** (**DD&AS**) is one of the components of the Blue-Cloud technical framework. It serves federated discovery and access to Blue Data Infrastructures (BDIs) for external users and also interacts with the Blue-Cloud Virtual Research Environment (VRE). As part of the predecessor pilot Blue-Cloud project, the first operational release of the Blue-Cloud DD&AS has been deployed. This service is federating in total 8 BDIs and with 9 data services. It provides a common user interface for discovery and retrieval of multi-disciplinary data sets as managed by each of the federated BDIs. Users are able to select and download selected data sets.

The implementation of the Blue Cloud DD&AS largely depends on machine-to-machine interactions between central components of the DD&AS and web services / APIs as managed and operated by the BDIs. The Blue-Cloud 2026 project strives for further optimisation of the DD&AS, in particular by analysing and where possible, improving the FAIRness of the services and contents as provided by the federated BDIs as well as by expanding the overall functionality of the DD&AS.

Therefore, the current federation situation per BDI as well as the functioning of central DD&AS components have been analysed and assessed for possible improvements in a cooperation and dialogue between the DD&AS core developers, MARIS and CNR-IIA, and managers and technicians of each of the BDIs. This analysis process will be described in this Deliverable D2.1, together with the proposed developments for making the improvements to the DD&AS.

Funded by the European Union
Blue-Cloud 2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227

2. Overall set-up of DD&AS

The first operational **Blue-Cloud Data Discovery and Access service** (DD&AS) is one of the two main components of the Blue-Cloud technical framework, next to the **Blue-Cloud Virtual Research Environment (VRE**), which has been designed and developed in the predecessor pilot Blue-Cloud project. The following paragraphs will describe this existing service.

2.1. Overall concept

The **Blue-Cloud Data Discovery and Access service** (DD&AS) facilitates discovery and retrieval of data sets and data products for external users in stand-alone mode, and for users of the VRE through connectivity. These data sets are managed in blue data infrastructures (BDIs) that are connected to the Blue-Cloud DD&AS to serve federated discovery and access.

The overall concept is that the Blue-Cloud DD&AS makes use of web services and APIs, following protocols such as CSW, OAI-PMH, ERDDAP, or otherwise, as provided and maintained by the BDIs. These are used to support machine-to-machine interactions for harvesting metadata, submitting queries, and retrieving resulting metadata, data sets, and data products.

The Blue-Cloud DD&AS provides a common interface for discovery and retrieval of data sets and data products from each of the federated BDIs, while the GUI also includes facilities for mapping and viewing the locations of data sets, as this is part of the query dialogue.

Conceptually, the query mechanism has a two-step approach:

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

The two-step approach for data discovery and access is effective to go from coarse to fine and to determine in an early stage which of the BDIs might have interesting data sets. It is also effective to keep the number of entries relatively limited in the exploratory first step of discovery. The granular level as a second level is applicable to several of the BDIs, in particular in cases with observation (raw) data which often can concern very large collections with numerous data sets. For instance, the SeaDataNet CDI

service currently gives at granular level discovery and access to more than 2.6 million individual observation data sets for physics, chemistry, geology, biology, geotechnics, and bathymetry. While at collection level, there are circa 900 aggregated CDI records, which lead to the circa 2.6 million granular records, which can be downloaded.

There are also cases, when one step is sufficient, such as in case of specific data products, that a user wants to download as a complete file. The second level then gives additional search criteria to discern better between data products and to allow the actual downloading.

In both cases, the overall principle is that queries at level 1 are applied to search on level 1 over all connected BDIs and on a common metadata profile, while queries at level 2 are applied with specific search profiles for each individual BD. For each BDI, selected level 2 records can be included in a shopping basket that could contain a mix of requested data sets from multiple BDIs. The shopping basket, once submitted, will deliver the requested data files in a data package, directly linked to the shopping order.

In the Blue-Cloud DD&AS for the first level use is made of the DAB metadata brokerage service software kit as developed and managed by CNR-IIA. The mappings are made against the common DAB metadata model, and the DAB service has been set up by CNR-IIA to generate, maintain, and provide a common Blue-Cloud level 1 catalogue as an internal service in a dynamical way with the latest entries as derived from the BDIs.

For the data access part of the Blue-Cloud DD&AS, a data brokerage service has been developed, integrating the internal Blue-Cloud level 1 metadata catalogue (see above), a series of machine-to-machine interfaces to the BDIs for level 2 queries, and a shopping mechanism to support the actual discovery and retrieval functions.

Implementing this approach fully depends on the interfaces of BDIs, that should be supportive for machine-to-machine interactions. For that reason, a lot of effort was dedicated to analysing and testing existing web services and APIs of each BDI for level 1, and to finding the best ways for the deployment of level 2 queries and how to establish the associated download URLs.

The β-version of the Blue-Cloud Data DD&AS has been launched in the middle of June 2021. Since then, further developments have taken place for adding functionality and refining specific parts of the service. This resulted in the current operational version of the DD&AS, which facilitates users:

- to search and discover interesting data sets;
- to complete and submit a shopping basket with interesting data sets;
- to stay informed about the progress of the shopping requests;
- to download the data sets once ready for downloading;
- to ingest data sets into the VRE data pool for use in VRE applications.

Furthermore, it facilitates managers of federated BDIs:

- to stay informed about the shopping requests and associated users for their repository;
- to prepare periodic management reports.

The operational DD&AS can be found at: https://data.blue-cloud.org/search

2.2. Architecture and components

The developments for the Blue-Cloud Data Discovery & Access Service have been led by MARIS, with contributions of CNR-IIA and EUDAT partners (DKRZ, CINECA, and CSC). MARIS has been responsible for the overall system and its integration, while CNR-IIA and EUDAT each have delivered components and contributed to the conceptual design, also involving IFREMER and CNR-ISTI. MARIS for a major part has made use of earlier developments and experiences gained during the upgrading of the SeaDataNet CDI service as part of the H2020 SeaDataCloud project. EUDAT has developed the Blue-Cloud data delivery component, also using the earlier achievements and experience built up during the H2020 SeaDataCloud project. Furthermore, there has been synergy with the recently finalised H2020 ENVRI-FAIR project, which reviewed and worked on improving FAIRness of data discovery and access services of multiple Research Infrastructures (RIs), active in 4 domains (marine, atmosphere, terrestrial, life). These experiences have contributed to the Blue-Cloud technical analyses.

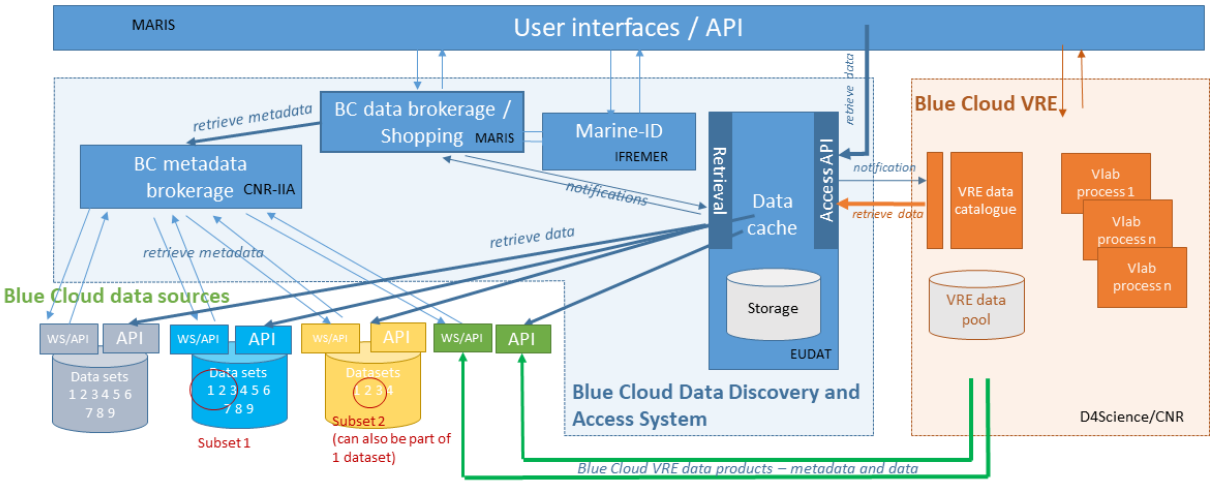

Figure 1 –Architecture of the Blue-Cloud discovery and access service

The figure 1 gives the architecture of the existing Blue-Cloud Data Discovery & Access service. It consists of a number of modules (services) as indicated in the figure:

- **Blue-Cloud metadata brokerage**, operated by CNR-IIA, dynamically interacting with each of the Blue Data Infrastructures (BDIs) to retrieve, extract and harmonise metadata entries for each BDI into a common Blue-Cloud level 1 metadata catalogue;
- **Blue-Cloud data sources**, comprising BDIs, that are gathering and managing catalogues and data collections from multiple observational platforms and/or from multiple data and data product originators. To illustrate the difference: Euro-Argo is an example of an observational network based upon Argo floats, deployed by multiple operators, and it manages all acquired metadata and data in a system with common discovery and access interfaces; SeaDataNet is a network of data centres, each serving one to many data providers for QA-QC and long term stewardship of their original datasets, while SeaDataNet operates an infrastructure for maintaining a central standardised catalogue and common discovery and access interfaces for data as managed and shared by its data centre nodes;
- **Marine-ID service,** operated by IFREMER, for registration and authentication of users to the Blue-Cloud Data Discovery and Access service. Users only have to register once to receive their login details;
- **Data cache**, operated by EUDAT, for temporary storage of data packages, consisting of data sets, retrieved from the Blue-Cloud data sources, plus associated metadata, as retrieved by the Blue-Cloud data brokerage. External users can download these data sets, after receiving information from the Blue-Cloud data brokerage, while the VRE can also be triggered to retrieve data packages for ingestion into the catalogue and data pool of the Blue-Cloud VRE;
- **Blue-Cloud Data brokerage service,** operated by MARIS. This service performs the master role in the Blue-Cloud Data Discovery and Access service, interacting with the other modules. Regularly (currently every week), it retrieves the latest Blue-Cloud level 1 metadata catalogue from the Blue-Cloud metadata brokerage, and ingests this into the discovery interface, whereby users can query the catalogue at level 1. The common level 1 metadata catalogue includes sufficient metadata for each BDI to allow the first level queries at collection level with a few selection criteria and this way to identify which of the BDIs holds interesting data sets. The Blue-Cloud level 1 metadata catalogue should also contain sufficient additional metadata to allow more specific searching at level 2 for those BDIs that only have data collections and other data products, but no service at granular level. While for other BDIs, supporting deeper searching at level 2 – granular level -, customised search profiles have been formulated, which allow the data broker to interact with the provided web services and APIs of the BDIs. The Blue-Cloud Data brokerage service also contains a shopping mechanism with basket and ledger, by which users (external users and VRE) and BDIs can be informed about shopping transactions and their status in time. It

Funded by the European Union
Blue-Cloud 2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227

interacts with the Blue-Cloud Data Cache to give it precise instructions about retrieving data sets from the BDIs and to insert these for temporary storage, and to bundle these as downloadable data packages for each shopping order. It interacts with the Marine-ID service as users need to login to submit shopping baskets and to have access to the transaction ledger. It interacts with registered users and VRE to inform and instruct them about data packages that are ready for downloading by users or retrieval for ingestion by the VRE. Finally, it also interacts with the Blue-Cloud Data Cache to receive information about the actual downloading by users and retrieval for ingestion by the VRE in order to update the ledger;

- **User interfaces,** operated by MARIS, to interact with users for discovery at level 1 and for deeper discovery and shopping transactions at level 2, and to provide access to the transaction ledger for users and BDIs.

Remark: In figure 1, the Blue-Cloud VRE is given with simplified graphics, only to underpin the exchange with the Blue-Cloud DD&AS.

In the following more detailed information is given about selected DD&AS components.

2.2.1 Blue-Cloud metadata brokerage

In the set-up of the Data Discovery & Access Service a distinction is made in metadata and data. The metadata are used for the discovery mechanism, building a common catalogue, while the data sets stay at the BDIs and can be retrieved in their existing formats from the BDIs, following the discovery process. For metadata the international ISO 19115 – 19139 metadata standard is used, which works with schema's and handles metadata in the machine-readable XML format. This choice is made because all BDIs are supporting webservices which facilitate to extract the metadata for their managed data sets. Having these then allows to derive a common denominator for the metadata of all federated BDIs. For the data level, the situation is quite different as between the BDIs there is a range of different data formats, also because the BDIs cover data from many disciplines which require different aspects to be included. For instance, NetCDF, ODV ASCII, other ASCII, and several other data formats are used, which sometimes also contain an overhead section with metadata next to the actual data. Converting all data sets to a common format, like e.g. NetCDF is not a valid option, considering the individual BDI operations, and also because there are a lot of software packages available and in use in the different communities that expect the specific data formats that the BDIs are supporting.

For the metadata brokerage at the first level of data collections, CNR-IIA has advanced and deployed an internal service, namely a Blue-Cloud discovery broker service based on their DAB technology. This middleware harvests metadata at collection level from each of the BDIs, using their existing web services or APIs. The DAB service transforms the harvested XML files from each of the BDIs into a common ISO

Blue-Cloud collection profile, which is published by the DAB service by means of a Blue-Cloud OGC-CSW service with a common XML profile for each BDI. See image below.

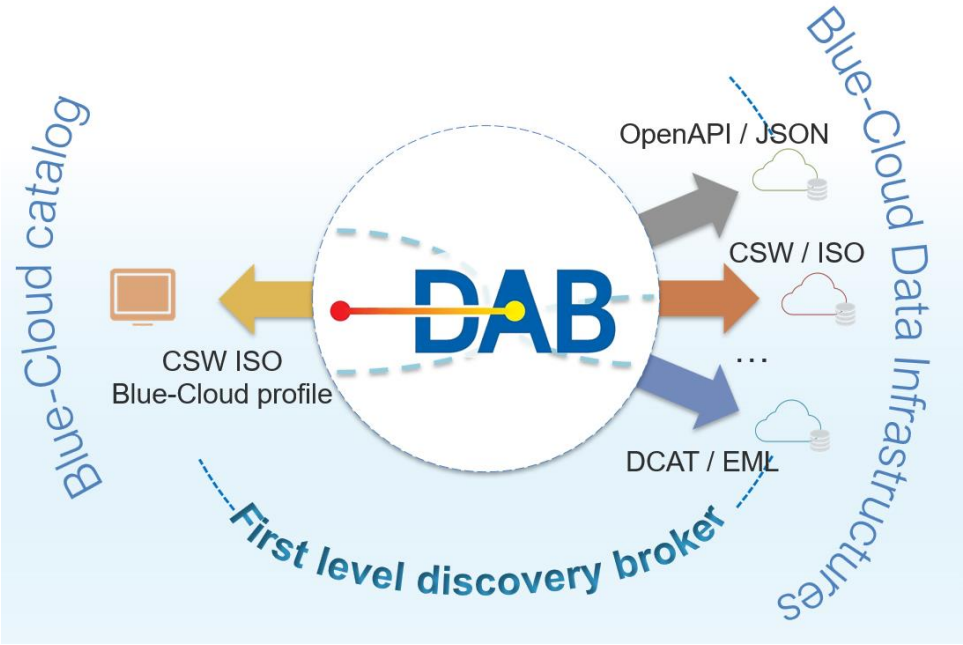

Figure 2 –Blue-Cloud first level discovery broker component harmonizes the protocols and data models published by different heterogeneous BDIs to a harmonized CSW service based on ISO 19115

The returned records are expressed according to the Blue-Cloud collection metadata profile, an ISO 19115 based metadata profile encoded using the recent ISO 19115-3:2016 XML schema implementation. In total 13 metadata elements from ISO 19115 are considered as the common elements of the profile, as they are deemed to be the more useful for discovery of Blue-Cloud collections. The common Blue-Cloud metadata elements are:

- IDENTIFIER: Blue-Cloud unique and persistent code for the metadata record;
- TITLE: a characteristic, and often unique, name by which the collection is known;
- ABSTRACT: a short description of the collection;
- KEYWORD: a commonly used word, formalised word or phrase used to describe the subject;
- BOUNDING_BOX: extent of the resource in the geographic space given as a bounding box;
- TEMPORAL_EXTENT: time period covered by the content of the collection;
- PARAMETER: name of the attribute described by the measurement value;
- INSTRUMENT: measuring instrument used to acquire the data;

Funded by the European Union
Blue-Cloud 2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227

- PLATFORM: platform from which the data were taken;
- ORGANIZATION: organization associated with the collection;
- DATESTAMP: the latest update date of the metadata description;
- REVISION_DATE: the latest update date of the data;
- RESOURCE_LINKS: download links where available and useful.

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

MARIS harvests these common formatted XML entries on a regular basis from each of the Blue-Cloud CSW services and integrates these into a SQL database which is indexed with Elastic Search, using Logstash as fast connector between the SQL database and the non-SQL Elastic Search database. This processing makes full free text searching very efficient and fast. Moreover, it facilitates building facets for level 2 queries for those BDIs which only have a collection / data products level. This way, the first step of the Blue-Cloud query process is powered, which has been integrated by MARIS in the interface of the Blue-Cloud Data DD&AS. And the common metadata base also serves the second level of the Blue-Cloud query process and the retrieval of data download links, but then only for the BDIs with one level of data. All has been set up as an automatic process without human intervention, driving weekly updating from the connected BDIs and synchronisation from the DAB CSW services to the indexed Blue-Cloud catalogue service as part of the Blue-Cloud DD&AS.

2.2.2 Blue-Cloud data brokerage and shopping mechanism

MARIS has integrated each BDI as part of the data brokerage service, arranging the 2nd query level and direct download links. The preferred way forward was to make use of web services and APIs at the BDIs to support building query profiles as a combination of facets and free search, providing results in a paging mode, and facilitating to browse detail pages per resulting record, and finally to retrieve the dedicated data links which are required for the shopping mechanism. Thereby, all these functions should be performed by automated and dynamic machine-to-machine interactions between the Blue-Cloud DD&AS and the web services or APIs of each BDI. In practice, deploying this preferred concept has not been feasible for each BDI as their web services were not (yet) fit and could not be adapted. In those cases, alternatives have been implemented, such as automatic direct harvesting of full metadata from a BDI and feeding these to the SQL db - Elastic Search db chain for building locally the facets, which then power the search. There is also quite some complexity, as there are several protocols being used, differing between the BDIs. This required that for each BDI a customised plug-in has been developed and deployed, also taking into account performance. In addition, hurdles have been overcome that some of the BDIs do not

provide direct machine-to-machine download links by https or ftp, but make use of HTML web forms. These had to be by-passed which required further communication with the BDI operators and adaptations from their side.

The shopping mechanism has been adopted and adapted using the earlier experience of MARIS and EUDAT partners with developing the SeaDataNet CDI service. It consists of:

- Shopping basket, which can be filled by users as a form, adding records from search results
- Marine-ID register, which holds account details for each user such as email, account name and password; shopping baskets can only be submitted by registered users;
- Transaction ledger, which holds information about all submitted shopping baskets and their status of processing;
- MyBlueCloud, dashboard for users to check order processing status and to download ready download packages;
- E-mails to users to confirm their submitted orders and to alert users when orders are completed and ready for downloading.

2.2.3 Blue-Cloud delivery component

EUDAT is in charge of the delivery service component of the Blue-Cloud Data Discovery and Access Service. This component is a temporary storage layer that is leveraged by the shopping basket interface to provide requested data to the final users. It exposes a dedicated API accessed from both the shopping basket interface for orders requests and from the final users for orders downloads. Once the user has filled the basket with requested data, the shopping system, run by MARIS, forwards the request to the data delivery component by providing a list of URLs in json format that are expected to be fetched from each of the BDIs.

Since the operation of downloading can involve potentially thousands of files, taking a long time, an asynchronous architecture is needed. For that reason, a shopping request is forwarded to a task worker that will download each file from the corresponding BDI onto a local filesystem. Once all files are downloaded, they are zipped together into one or more archives to ease the subsequent download. Multiple archives can be created to prevent the creation of final files larger than a threshold (< 2 GB). The shopping basket is notified when the download process is completed and the order is ready for the download. By contacting the API again, the shopping basket will retrieve a list of URLs that will be provided to the user through its MyBlueCloud dashboard for actual downloading. The URL embeds an identification token to authorize the download, which ensures that download URLs can only be obtained from the shopping basket interface to prevent any data leakage.

Data Cache API endpoints are written by EUDAT in Python by adopting the Flask microframework and served through a nginx reverse proxy, also providing a HTTP over TLS connection. Both user credentials and session tokens are stored into a dedicated PostgreSQL database. Asynchronous jobs are implemented in Python by using the Celery task queue, backed with a results database implemented with Redis. Redis is also responsible for the communication between Flask APIs and the Celery tasks by operating as message broker system.

Retrieved data can be accessed by two different paths. Next to the direct download by users (see description earlier), it is also possible to enable transfer to the VRE data pool. The latter is required as VRE users might want to use the downloaded data for further processing on the powerful D4Science VRE premises. This could be for providing input to Virtual Labs developments and also for the newly planned Work Benches developments. The VRE transfer takes place in a very similar way as the download by users described above, except that the download will be carried out by the VRE. The experiences in SeaDataCloud have shown that pulling/downloading data is much more efficient and less error-prone than pushing/uploading the data, so the implemented workflow is:

- The EUDAT API receives a request for transferring a batch of data to the VRE data pool (request from data broker to EUDAT API);
- The EUDAT component provides to the data broker the URLs to transfer the requested data batch (asynchronous request from EUDAT to data broker);
- The data broker notifies to the VRE that the data batch is available for pull (asynchronous request from data broker to VRE API);
- The VRE downloads the data using the specified link (request from VRE to EUDAT API).

Data are maintained into the Data Cache based on normal data retention policies (30 days). To realize this workflow a specific component has been developed and added by CNR-ISTI to the VRE to handle requests to pull data. As both CNR-ISTI and EUDAT are part of the European research network GÉANT, this transfer benefits from the very fast network backbone between major European research centres. The Data cache API is currently deployed on the HPC cloud infrastructure at CINECA (ADA cloud), based on OpenStack Wallaby and equipped with:

- 68 interactive OpenStack nodes each 2 x CPU Intel CascadeLake 8260, with 24 cores each, 2,4 GHz, 768GB RAM and 2TB SSD storage;
- 1 PB Ceph storage raw dedicated (full NVMe/SSD);

This cloud infrastructure is tightly connected both to the LUSTRE storage of 20 PB raw capacity, and to the GSS storage of 6 PB seen by all other infrastructure.

2.2.4 Blue-Cloud DD&AS User Interface

The Blue-Cloud Data DD&AS provides a common interface for discovery and retrieval of data collections from the federated BDIs. The GUI also includes facilities for mapping and viewing the locations of data sets, as this will be part of the query dialogue. Moreover, the GUI includes the query mechanism as a two-step approach, whereby the first step has a focus on identifying interesting data collections and products, while the second step facilitates drilling down and subsetting within the identified collections and products in order to get more specific data. For the second step, geographic and temporal criteria are instrumental, next to additional criteria which are specific per BDI. Finally, users are able to download and store the retrieved data collections on their own machines or in a data pool as part of the Blue-Cloud VRE.

The following screens illustrate the current User Interface dialogues.

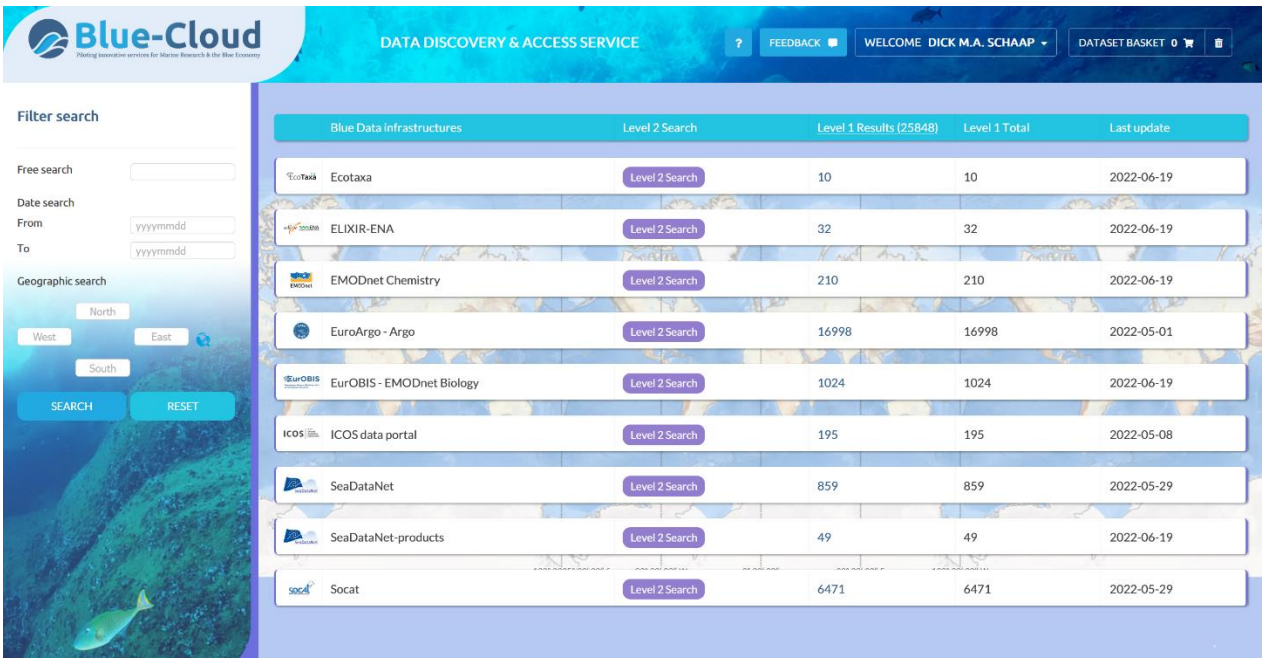

Figure 3 –Impression of Blue-Cloud Data Discovery & Access service – 1st level search on data

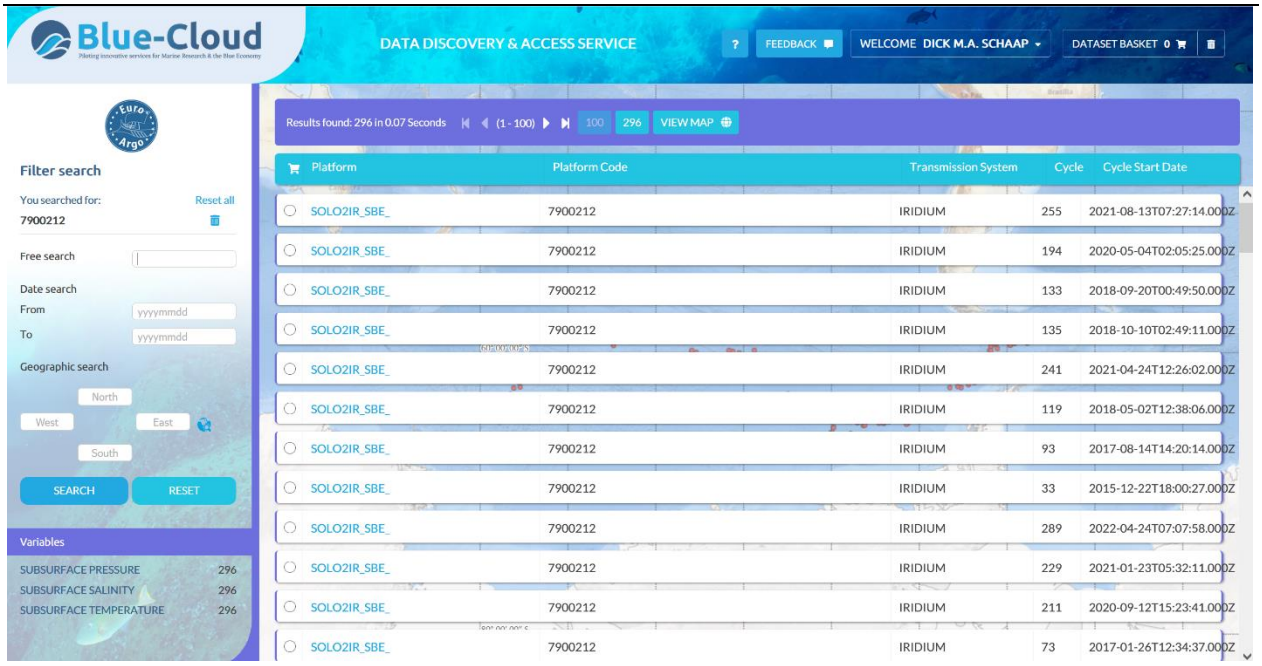

Figure 4 –Impression of Blue-Cloud Data Discovery & Access service – 2nd level search on Argo data individual records

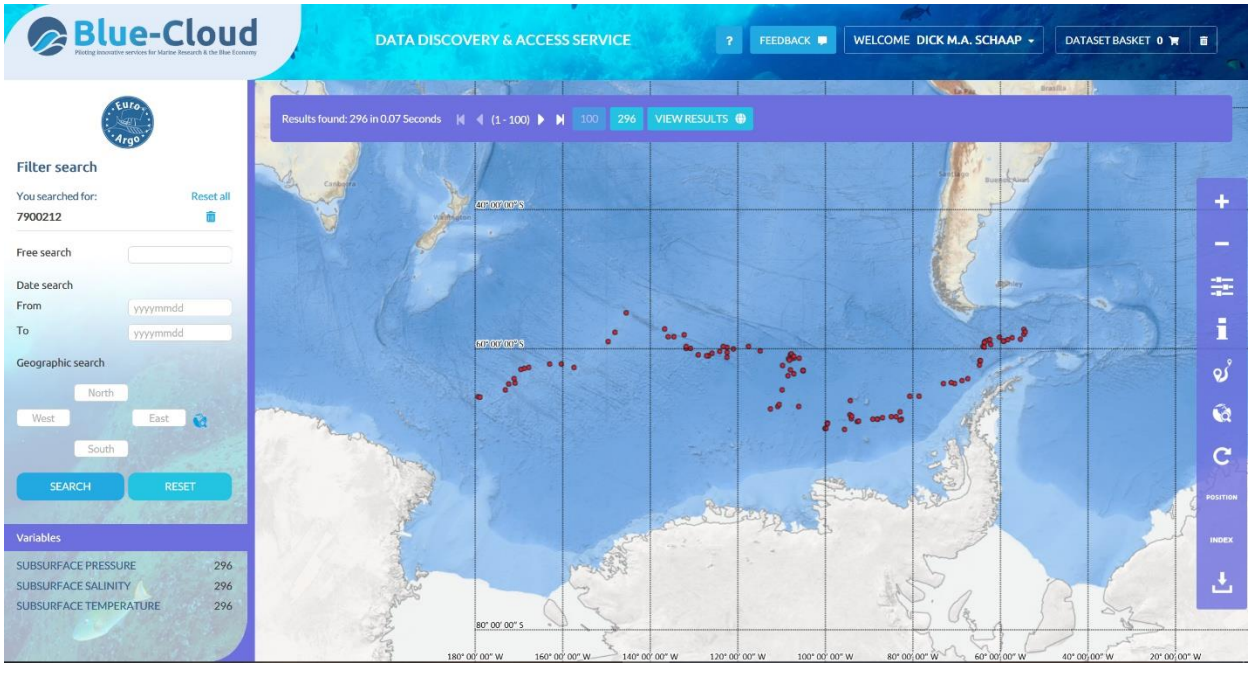

Figure 5 – Impression of Blue-Cloud Data Discovery & Access service – 2nd level search and Map on Argo data individual records

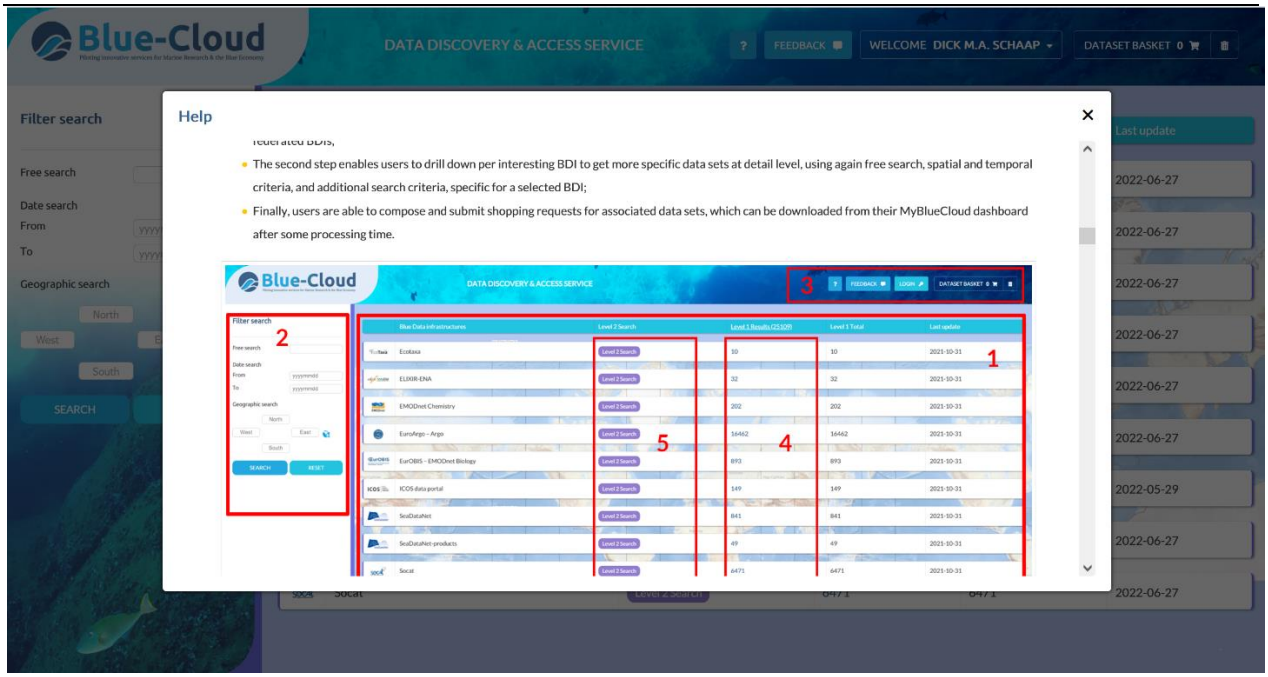

Figure 6 - Impression of the Help section on the Blue-Cloud Data Discovery & Access service

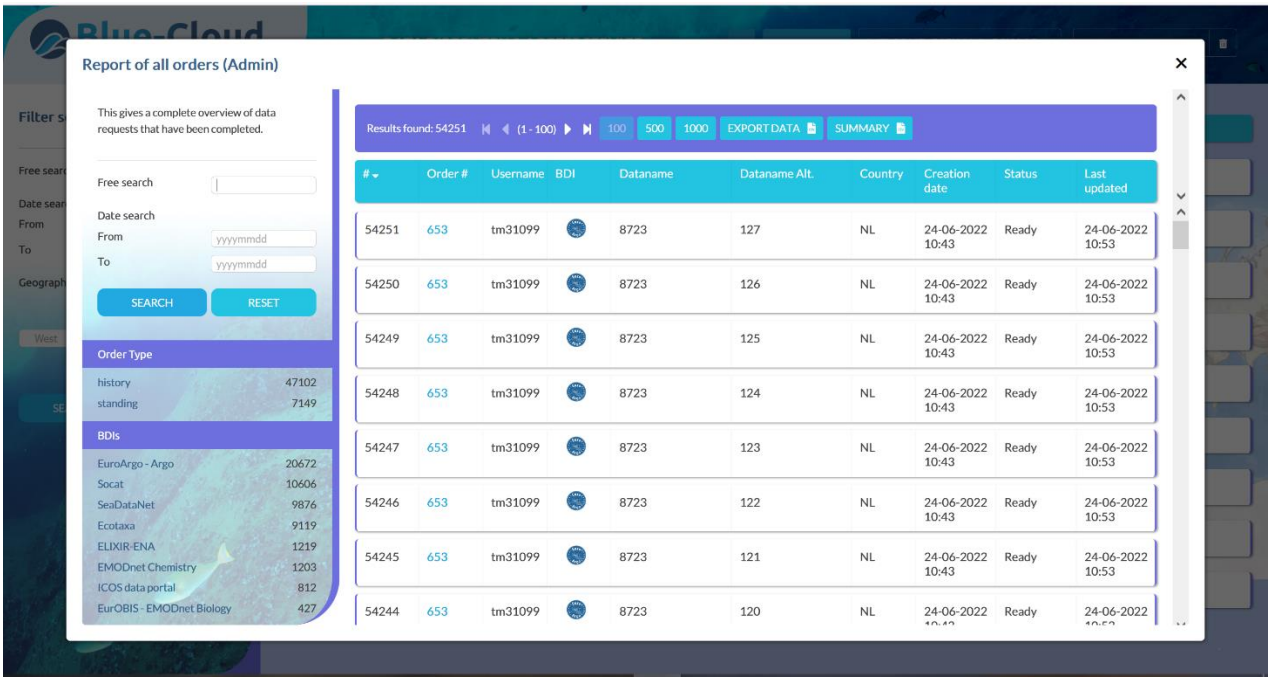

Figure 7 - Impression of the Report of all Orders section on the Blue-Cloud Data Discovery & Access service

3. FAIRness of services

3.1. How to assess FAIRness of services

One of Blue-Cloud's core aims is to offer users Virtual Labs, Work Benches and a VRE in which they can process and visualise marine datasets from a wide range of Blue Data Infrastructures. To streamline the access to data Blue-Cloud has developed the Data Discovery and Access Service (DD&AS) which offers in two levels and with broker involvement access to the datasets published by the BDI's. When analysing the existing DD&AS and Blue Data Infrastructures and the work done to create access to the data, we find that - although there is more convergence in the published metadata at collection level (via services like CSW, OAI-PMH and in standards like ISO 19115/19139) - there is a large diversity in the actual data access services. Even if a service is machine-to-machine accessible, there might be issues like a) the documentation is difficult to find, b) it does not follow standards, c) output and input are not clearly defined. In the development of the DD&AS the broker software largely overcomes this hurdle, but for the longer term it would be more efficient (for maintenance and also in relation to usage in other systems) if the services would work towards improved FAIRness.

For the optimization of the Blue-Cloud DD&AS it is not only important to focus on the FAIRness of the digital objects, such as the metadata, data, and data products, that are being managed and offered by the BDIs. The FAIRness optimization also needs to focus on the infrastructures and the services at the BDIs that are providing discovery and access functionalities for those digital objects. Existing FAIRness assessments methods, such as for instance the **GO FAIR Convergence Matrix**[1] as used earlier in the ENVRI-FAIR project[2] for reviewing several BDIs[3], mostly focuses on evaluating the FAIRness of the digital objects, while hardly evaluating the underpinning services.

Therefore, it would be useful if use could be made of an assessment framework for data services. However, such assessment frameworks are scarce, which motivated the FAIRsFAIR project[4] - Fostering FAIR Data Practices in Europe - to initiate formulating such a framework. Their work was inspired by a combination of literature describing the expectations users have from FAIR data services, and refined by the authors based on feedback from the community gained e.g. through workshops.

[1] *https://www.go-fair.org/today/FAIR-matrix/*
[2] *https://envri.eu/home-envri-fair/*
[3] *https://zenodo.org/record/388529*
[4] *https://www.fairsfair.eu/*

The resulting **Framework for assessing FAIR Services** focuses on providing guidelines on how services can be made to optimally improve the FAIRness of the data that they are used for. This work was inspired by a combination of literature describing the expectations users have from FAIR data services, and refined by the authors based on feedback.

The FAIRsFAIR service assessment framework includes seven aspects and fifty recommendations at different priority levels. The following table lists the seven aspects, while the full framework can be found in the report[5].

Aspect	Objective
FAIR enablement	The service enables FAIR data by elevating the FAIRness of digital objects and/or supporting the FAIRification process. FAIR enablement is actively driven through the implementation of community-supported standards and interoperability frameworks.
Quality of Service	The service is delivered in a reliable, secure, high-quality way, consistent with its specifications
Open & Connected	The service is operated in a low-barrier and inclusive way, seeking integrations and connections with other services and championing principles of openness consistent with Open Science and Open Research.
User Centricity	The service is managed so that it serves the (possibly evolving) goals of the user community and maximises usability while minimizing burden.
Transparency	The service provider communicates with its stakeholders in a transparent manner.
Longevity	The service provider designs the service with a timeframe for the maintenance and sustainability of the service in mind and implements measures accordingly, considering the researchers' need for reproducible research.
Ethical & Legal	The service complies with all applicable legal and ethical guidelines, in a transparent and auditable way.

Table 1: Aspects of the FAIRsFAIR Framework for assessing FAIR Services

This framework specifically targets service owners (i.e. individuals or organisations developing and operating services for research data). The framework does not assign a FAIRness score to a service, its purpose is to support services in becoming more FAIR-enabling by highlighting areas for improvement.

[5] https://zenodo.org/record/6656431

ACTION: **WP2 will organize a workshop early 2024 as part of the Technical and Scientific Committee (TSC) to make together with representatives of all BDIs an assessment of their data discovery and access services following the aspects and recommendations of the FAIRsFAIR Framework for assessing FAIR Services. It will be tried to involve author(s) of the framework as guides as it is important to have a common interpretation of specific recommendations and their perspectives.**

This 'holistic' assessment will provide insights and suggestions for possible improvements that could be made to increase the FAIRness of the BDI services, considering the broader spectrum of aspects, such as organisational, documentation, and legal aspects.

3.2 Documenting services in a FAIR way

Documenting services is one of the recommendations in the FAIRness framework and is essential for lowering the threshold for being able to understand the functionality of services, what input and what output can be expected, what processing capacity is required, what is the licence for use etc. "Just providing APIs" is not sufficient. In the project FAIR-IMPACT[6] and RDA[7] working groups there has been developed a basis for documenting FAIR software including validation. Also in schema.org a basis is available for describing services. But both are not yet sufficient and need further work. As a basis each data discovery and access service will need a FAIR service description, partly similar to FAIR software.

This aspect of FAIR documenting services is also being explored as part of the FAIR-EASE project[8].

ACTION : Blue-Cloud will work together with projects like FAIR-EASE and FAIR-IMPACT on developing an optimised FAIR service description model that then could be adopted by each of the BDIs for a local implementation.

[6] *https://fair-impact.eu/*
[7] *https://www.rd-alliance.org/*
[8] *https://fairease.eu/*

4. Situation and technical analysis per BDI

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

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

These principles have been taken onboard by the DD&AS core developers for reviewing the current federation deployments and in the bilateral meetings between the core developers and managers and technicians of each of the BDIs. These bilateral meetings took place in the period from August to October 2023 and were dedicated to discuss findings and to brainstorm about suggestions and options for optimising the technical federations.

The following Blue Data Infrastructures (BDIs) have been federated in the current operational release of the Blue-Cloud Data Discovery & Access service:

- EMODnet Chemistry Products service
- SeaDataNet Data Products service
- SeaDataNet CDI Data service (including also CDI entries for EMODnet Chemistry, Bathymetry, and Physics)
- EurOBIS – EMODnet Biology data collections service
- Euro-Argo - Argo GDAC data services
- ELIXIR-ENA data service
- EcoTaxa data service
- ICOS-Marine data service
- ICOS-SOCAT data service

In the following paragraphs, for each BDI a description will be given about the infrastructure, how they are currently federated at level 1 and level 2 in the DD&AS, and what optimisations could be undertaken, both by the BDIs and the DD&AS core services. These are formulated as actions for the coming developments.

4.1. EMODnet Chemistry Products service

The European Marine Observation and Data network (EMODnet) [9] 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 themes 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 these 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. Next to the seven thematic groups, there is also EMODnet Ingestion which encourages and facilitates additional data managers to ingest their marine datasets for further processing, publishing as open data and contributing to applications for society. Early 2023, the earlier seven thematic portals have been migrated and integrated into one single shared EMODnet portal,

[9] *https://emodnet.ec.europa.eu/*

which provides a number of central services, such as a central map viewing service and a central products catalogue service, which are combining input from all seven thematic groups. This is done by machine-to-machine services. The central EMODnet portal operates the front-ends of the central services, while thematic groups operate back-ends with OGC services for exchanges.

4.1.1 Description

The EMODnet Chemistry partnership 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 expertise and experiences of collecting, processing, and managing of chemistry data together with expertise 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 gathering is done by using the SeaDataNet CDI service infrastructure. The gathered data are then aggregated and validated by MSFD region. 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. 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. The collected marine litter data are then centrally converted to entries for the SeaDataNet CDI service. While for micro

plastics the regular SeaDataNet CDI service approach has been adapted and dedicated guidelines have been formulated and published in concertation with the TG-ML.

All resulting products for eutrophication, contaminants, and marine litter are described with metadata in the local Chemistry products catalogue, have DOIs and landing pages for citation, and are also made available by OGC (CSW, WMS, WFS) web services for inclusion in the central EMODnet Products Catalogue and the central EMODnet Map Viewer service.

4.1.2 Current federation

The current federation has been directed to the EMODnet Chemistry products catalogue which publishes Chemistry products through a OGC CSW/ISO v.2.0.2 interface at:

https://sextant.ifremer.fr/geonetwork/EMODNET_Chemistry/eng/csw?service=CSW&request=GetCapabilities&version=2.0.2

Currently, more than 300 records are discoverable from the service, each one is described with an ISO 19115 metadata document.

The OGC CSW service is a standard protocol which makes it easy for the Blue-Cloud DAB metadata broker service to harvest and extract the fields as specified for the common profile. (see paragraph 2.2.1).

The EMODnet Chemistry standard CSW ISO interface allows to retrieve metadata records executing HTTP-GET GetCapabilities operation followed by multiple paginated HTTP-POST GetRecords operations to harvest all the catalog content. Each MD_Metadata record is mapped to a Blue-Cloud collection record.

In the following table the mapping towards the Blue-Cloud metadata core elements is reported.

Blue-Cloud core metadata element	EMODnet Chemistry metadata element
Identifier	/gmd:MD_Metadata/gmd:fileIdentifier
Title	/gmd:MD_Metadata/ gmd:identificationInfo/gmd:MD_DataIdentification/gmd:citation/gmd:CI_Citation/gmd:title
Keyword	/gmd:MD_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification /gmd:descriptiveKeywords/gmd:MD_Keywords/gmd:keyword
Bounding box	/gmd:MD_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification /gmd:extent/gmd:EX_Extent/gmd:geographicElement/gmd:EX_GeographicBoundingBox

Blue-Cloud core metadata element	EMODnet Chemistry metadata element
Temporal extent	/gmd:MD_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent /gmd:EX_Extent/gmd:temporalElement/gmd:EX_TemporalExtent/gmd:extent
Parameter	/gmd:MD_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords[gmd:type/gmd:MD_KeywordTypeCode/@codeListValue='parameter']/gmd:keyword/
Instrument	Not available for this BDI
Platform	Not available for this BDI
Organization	/gmd:MD_Metadata/gmd: identificationInfo /gmd:MD_DataIdentification/gmd:pointOfContact/gmd: CI_ResponsibleParty/gmd:organisationName
Date stamp	/gmd:MD_Metadata/gmd:dateStamp
Revision date	Seems to be not available for this BDI

Table 2: mapping of Chemistry metadata elements to Blue-Cloud core elements

Level 1:

As a result, the DAB broker publishes EMODnet Chemistry products through the following endpoint:

https://blue-cloud.geodab.eu/gs-service/services/essi/view/emodnet-chemistry/csw

with following GetCapabilities

https://blue-cloud.geodab.eu/gs-service/services/essi/view/emodnet-chemistry/csw?service=CSW&request=GetCapabilities&version=2.0.2

The returned records are expressed according to the Blue-Cloud metadata profile, that is a ISO 19115 based metadata profile encoded using the recent ISO 19115-3:2016 XML schema.

To assess the quantity and quality of the overall Blue-Cloud metadata content a metadata report is made available at:

https://dabreporting.s3.amazonaws.com/BlueCloud/BlueCloudReport_brief.html

The report shows with graphical indicators the most recent status of the Blue-Cloud metadata content.

Funded by the European Union
Blue-Cloud 2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227

EMODnet Chemistry available service at: **https://blue-cloud.geodab.eu/gs-service/services/essi/view/emodnet-chemistry/csw**
EMODnet Chemistry available test portal at: https://blue-cloud.geodab.eu/gs-service/search?view=emodnet-chemistry
Total number of records: **345** Number of records analyzed: **345** Percentage of records analyzed: **100%**

Metadata element	Path	Completeness
IDENTIFIER	//gmd:fileIdentifier/gco:CharacterString	100%
TITLE	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:citation/gmd:CI_Citation/gmd:title/*[1]	96%
KEYWORD	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords[not(gmd:type) or not(contains('platform instrument',gmd:type/gmd:MD_KeywordTypeCode/@codeListValue))]/gmd:keyword/*[1]	64%
BOUNDING_BOX	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:geographicElement /gmd:EX_GeographicBoundingBox/gmd:westBoundLongitude/gco:Decimal /gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:geographicElement /gmd:EX_GeographicBoundingBox/gmd:eastBoundLongitude/gco:Decimal /gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:geographicElement /gmd:EX_GeographicBoundingBox/gmd:southBoundLatitude/gco:Decimal /gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:geographicElement /gmd:EX_GeographicBoundingBox/gmd:northBoundLatitude/gco:Decimal	96%
TEMPORAL_EXTENT	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:temporalElement/gmd:EX_TemporalExtent /gmd:extent/gml32:TimePeriod/gml32:beginPosition /gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:temporalElement/gmd:EX_TemporalExtent /gmd:extent/gml32:TimePeriod/gml32:endPosition	93%
PARAMETER	/gmi2019:MI_Metadata/gmd:contentInfo/gmi2019:MI_CoverageDescription/gmd:attributeDescription/gco:RecordType	96%
INSTRUMENT	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords[gmd:type /gmd:MD_KeywordTypeCode/@codeListValue='instrument']/gmd:keyword/*[1]	0%
PLATFORM	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords[gmd:type /gmd:MD_KeywordTypeCode/@codeListValue='platform']/gmd:keyword/*[1]	0%
ORGANIZATION	//gmd:CI_ResponsibleParty/gmd:organisationName/*[1]	99%
DATESTAMP	/gmi2019:MI_Metadata/gmd:dateStamp/gco:Date	100%
REVISION_DATE	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:citation/gmd:CI_Citation/gmd:date/gmd:CI_Date[gmd:dateType /gmd:CI_DateTypeCode/@codeListValue='revision']/gmd:date/gco:Date	0%
RESOURCE_IDENTIFIER	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:citation/gmd:CI_Citation/gmd:identifier/gmd:MD_Identifier/gmd:code /gco:CharacterString	96%

Figure 8 DAB completeness report of core metadata elements in the EMODnet Chemistry products service as determined from their web service

The output of the Blue-Cloud DAB broker is directly used to drive the level 1 discovery service of the Blue-Cloud Data Discovery & Access service. Therefore, a filter has been included to ensure that all included records have a download option.

Level 2:

In this case, there is only a collections level and therefore the level search and actual downloading is also derived from the Blue-Cloud DAB broker service output. For the level 2 search extra search criteria from the common profile are used:

- Free search
- Lat-Lon box
- Date period
- Keywords
- Parameters
- Organisation

4.1.3 Proposed optimisation

At the start of the analysis it appeared that only a reduced number of harvested and converted records came through the filter checking for direct downloads. It appeared that for many records in the Chemistry Catalogue there were issues with direct download links. This was a side effect from the earlier migration to the central EMODnet portal, which was concluded early 2023, and which required a reorganisation of the download services. Together with the EMODnet Chemistry technical team, this issue has been reviewed and cleared, resulting in a considerable larger number of EMODnet Chemistry products that now can be downloaded and which are thus made available in the Blue-Cloud DD&AS.

The Chemistry Sextant service itself is hosted at Ifremer as part of EMODnet Chemistry and is a sustained service as part of the EMODnet Initiative, funded by EU DG MARE since 2008 and with long term perspective.

Another aspect is semantic interoperability. For the population of the EMODnet Chemistry Sextant catalogue, hosted as GeoNetwork catalogue, use is made of controlled vocabularies of SeaDataNet for fields such as parameter (P02) and organisations (EDMO). However, when harvesting and extracting into the Blue-Cloud DAB service, these vocabulary references such as code and underlying vocabulary service are not coming back. The terms only come back with their controlled descriptions. One of the targets of the optimisation of the DD&AS is to add a semantic brokerage to the DD&AS. For this it will be very helpful if the metadata is enriched with the codes and associated vocabulary services as that might facilitate recognising vocabularies and automated mapping between terms originating from different Blue Data Infrastructures.

4.1.4 Planned actions

ACTION : Analyse how the coding and associated vocabularies as used at the EMODnet Chemistry catalogue source could be added and included in the common profile of the Blue-Cloud DAB service and served out through the DAB OGC CSW service endpoint for EMODnet Chemistry Products in order to facilitate semantic brokering. This action will be undertaken by CNR-IIA and MARIS.

4.2. SeaDataNet Data Products service

SeaDataNet[10] 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.

4.2.1 Description

SeaDataNet provides aggregated datasets (ODV collections of all unrestricted SeaDataNet measurements of temperature and salinity by sea basins) and climatologies (regional gridded field products) based on the aggregated datasets and data from external data sources such as the COriolis Ocean Dataset for Reanalysis and the World Ocean Database (WOD) for all the European sea basins and the Global Ocean. Each SeaDataCloud product is described in the SeaDataNet Products catalogue service.

4.2.2 Current federation

The SeaDataNet products catalogue publishes SeaDataNet products through a OGC CSW/ISO v.2.0.2 interface at:

https://sextant.ifremer.fr/geonetwork/SEADATANET/eng/csw?service=CSW&request=GetCapabilities&version=2.0.2

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

Currently, more than 30 records are discoverable from the service, each one is described with an ISO 19115 metadata document.

The OGC CSW service is a standard protocol which makes it easy for the Blue-Cloud DAB metadata broker service to harvest and extract the fields as specified for the common profile. (see paragraph 2.2.1).

The SeaDataNet Products standard CSW ISO interface allows to retrieve metadata records executing HTTP-GET GetCapabilities operation followed by multiple paginated HTTP-POST GetRecords operations to harvest all the catalog content. Each MD_Metadata record is mapped to a Blue-Cloud collection record.

In the following table the mapping towards the Blue-Cloud metadata core elements is reported.

Blue-Cloud core metadata element	SeaDataNet Products metadata element
Identifier	/gmd:MD_Metadata/gmd:fileIdentifier
Title	/gmd:MD_Metadata/ gmd:identificationInfo/gmd:MD_DataIdentification/gmd:citation/gmd:CI_Citation/gmd:title
Keyword	/gmd:MD_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification /gmd:descriptiveKeywords/gmd:MD_Keywords/gmd:keyword
Bounding box	/gmd:MD_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification /gmd:extent/gmd:EX_Extent/gmd:geographicElement/gmd:EX_GeographicBoundingBox
Temporal extent	/gmd:MD_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent /gmd:EX_Extent/gmd:temporalElement/gmd:EX_TemporalExtent/gmd:extent
Parameter	/gmd:MD_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords[gmd:type/gmd:MD_KeywordTypeCode/@codeListValue='parameter']/gmd:keyword/
Instrument	Not available for this BDI
Platform	Not available for this BDI
Organization	/gmd:MD_Metadata/gmd: identificationInfo /gmd:MD_DataIdentification/gmd:pointOfContact/gmd:CI_ResponsibleParty/gmd:organisationName
Date stamp	/gmd:MD_Metadata/gmd:dateStamp
Revision date	/gmd:MD_Metadata/gmd:identificationInfo/sdn:SDN_DataIdentification/gmd:citation/gmd:CI_Citation/gmd:date/gmd:CI_D

Blue-Cloud core metadata element	SeaDataNet Products metadata element
	ate[gmd:dateType/gmd:CI_DateTypeCode/@codeListValue='revision']/gmd:date

Table 3: mapping of SeaDataNet Products metadata elements to Blue-Cloud core elements

Level 1:

As a result, the DAB broker publishes SeaDataNet Products through the following endpoint:

https://blue-cloud.geodab.eu/gs-service/services/essi/view/seadatanet-products/csw

with following GetCapabilities

https://blue-cloud.geodab.eu/gs-service/services/essi/view/seadatanet-products/csw?service=CSW&request=GetCapabilities&version=2.0.2

The returned records are expressed according to the Blue-Cloud metadata profile, that is a ISO 19115 based metadata profile encoded using the recent ISO 19115-3:2016 XML schema.

To assess the quantity and quality of the overall Blue-Cloud metadata content a metadata report is made available at:

https://dabreporting.s3.amazonaws.com/BlueCloud/BlueCloudReport_brief.html

The report shows with graphical indicators the most recent status of the Blue-Cloud metadata content.

Funded by the European Union
Blue-Cloud 2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227

SeaDataNet Products available service at: **https://blue-cloud.geodab.eu/gs-service/services/essi/view/seadatanet-products/csw**
SeaDataNet Products available test portal at: https://blue-cloud.geodab.eu/gs-service/search?view=seadatanet-products
Total number of records: **51** Number of records analyzed: **51** Percentage of records analyzed: **100%**

Metadata element	Path	Completeness
IDENTIFIER	//gmd:fileIdentifier/gco:CharacterString	100%
TITLE	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:citation/gmd:CI_Citation/gmd:title/*[1]	100%
KEYWORD	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords[not(gmd:type) or not(contains('platform instrument',gmd:type/gmd:MD_KeywordTypeCode/@codeListValue))]/gmd:keyword/*[1]	100%
BOUNDING_BOX	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:geographicElement /gmd:EX_GeographicBoundingBox/gmd:westBoundLongitude/gco:Decimal /gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:geographicElement /gmd:EX_GeographicBoundingBox/gmd:eastBoundLongitude/gco:Decimal /gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:geographicElement /gmd:EX_GeographicBoundingBox/gmd:southBoundLatitude/gco:Decimal /gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:geographicElement /gmd:EX_GeographicBoundingBox/gmd:northBoundLatitude/gco:Decimal	100%
TEMPORAL_EXTENT	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:temporalElement/gmd:EX_TemporalExtent /gmd:extent/gml32:TimePeriod/gml32:beginPosition /gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:temporalElement/gmd:EX_TemporalExtent /gmd:extent/gml32:TimePeriod/gml32:endPosition	100%
PARAMETER	/gmi2019:MI_Metadata/gmd:contentInfo/gmi2019:MI_CoverageDescription/gmd:attributeDescription/gco:RecordType	96%
INSTRUMENT	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords[gmd:type /gmd:MD_KeywordTypeCode/@codeListValue='instrument']/gmd:keyword/*[1]	0%
PLATFORM	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords[gmd:type /gmd:MD_KeywordTypeCode/@codeListValue='platform']/gmd:keyword/*[1]	0%
ORGANIZATION	//gmd:CI_ResponsibleParty/gmd:organisationName/*[1]	100%
DATESTAMP	/gmi2019:MI_Metadata/gmd:dateStamp/gco:Date	100%
REVISION_DATE	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:citation/gmd:CI_Citation/gmd:date/gmd:CI_Date[gmd:dateType /gmd:CI_DateTypeCode/@codeListValue='revision']/gmd:date/gco:Date	0%
RESOURCE_IDENTIFIER	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:citation/gmd:CI_Citation/gmd:identifier/gmd:MD_Identifier/gmd:code /gco:CharacterString	100%

Figure 9 DAB completeness report of core metadata elements in the SeaDataNet Products service as determined from their web service

The output of the Blue-Cloud DAB broker is directly used to drive the level 1 discovery service of the Blue-Cloud Data Discovery & Access service. Therefore, a filter has been included to ensure that all included records have a download option.

Level 2:

In this case, there is only a collections level and therefore the level search and actual downloading is also derived from the Blue-Cloud DAB broker service output. For the level 2 search extra search criteria from the common profile are used, next to the universal criteria of level 1:

- Free search
- Lat-Lon box
- Date period
- Keywords
- Parameters
- Organisation

Funded by the European Union
Blue-Cloud 2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227

4.2.3 Proposed optimisation

The SeaDataNet Products catalogue service is stable as it contains products that have been generated in the EU projects for SeaDataNet and SeaDataCloud. No recent products have been added. The service itself is hosted at Ifremer as part of the SeaDataNet AISBL as a sustained service and is continuosly monitored for availability with excellent results.

The focus for optimisation in Blue-Cloud 2026 will be on semantic interoperability. The SeaDataNet Products catalogue is hosted as GeoNetwork catalogue by Ifremer as part of the Sextant Catalogue service which is also hosting the EMODnet Chemistry Products catalogue. In fact, the two catalogues are very similar in their set-up. So, likewise, it will be very helpful if the output metadata of the SeaDataNet Products Catalogue is enriched with the codes and associated vocabulary services as that might facilitate recognising vocabularies and automated mapping between terms originating from different Blue Data Infrastructures.

4.2.4 Planned actions

ACTION : Analyse how the coding and associated vocabularies as used at the SeaDataNet Products catalogue source could be added and included in the common profile of the Blue-Cloud DAB service and served out through the DAB OGC CSW service endpoint for SeaDataNet Products in order to facilitate semantic brokering. This action will be undertaken by CNR-IIA and MARIS and is combined with the earlier action for EMODnet Chemistry products.

4.3. SeaDataNet CDI Data service

See the previous paragraph 4.2 for a general SeaDataNet description. Next to the SeaDataNet Data Products catalogue service, SeaDataNet maintains a range of pan-European directories, such as for projects (EDMERP), organisations (EDMO), large data sets (EDMED), cruise summary reports (CSR), and monitoring programs and stations (EDIOS), and a wide range of controlled vocabularies (SeaDataNet vocabularies). All these directories and vocabularies are published as User Interfaces and as web services (SOAP – SPARQL). A core service is the SeaDataNet CDI service, which is federated as part of Blue-Cloud DD&AS.

4.3.1 Description

The **SeaDataNet Common Data Index (CDI) data discovery and access service** 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. 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 **2.8 Million data sets**, originating from more than **900 organisations** in Europe, covering physical, geological, chemical, biological and geophysical data, and acquired in European waters and global oceans.

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. Most data are available with CC-BY-4.0 license and are automatically delivered after requests from the SeaDataNet cloud storage for public data. The User Interface also has means for requesting access to restricted data, but for the Blue-Cloud DD&AS the focus is on the open data, which are currently more than 2.6 million data sets. The CDI metadata format is a marine profile of the ISO 19115 – 19139 standard, is fully supported by the SeaDataNet controlled vocabularies and directories (EDMERP, EDMED, EDMO) and is INSPIRE compliant. Most of the data sets have a common data format such as SeaDataNet ODV ASCII and NetCDF, except for data types that come with their own standards, such as e.g. SEG-Y for seismics, XYZ, BAG, and GeoTiff for bathymetry, and ASCII formats for marine litter types.

4.3.2 Current federation

For the CDI service, SeaDataNet operates a web service at granular level and also one at collection level. The latter is done automatically by aggregation of CDIs:

- 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 2.6 million open CDI records result in circa 900+ aggregated CDI records, which make use of the same CDI INSPIRE compliant metadata schema.

The URL of the CDI aggregated web service is:

https://cdi.seadatanet.org/report/aggregation/open

−<cdiGroup>
−<cdiUrl>
https://cdi.seadatanet.org/report/aggregation/486/486/486/4/ds03/open/xml
</cdiUrl>
−<cdiUrl>
https://cdi.seadatanet.org/report/aggregation/486/486/486/4/ds07/open/xml
</cdiUrl>
−<cdiUrl>
https://cdi.seadatanet.org/report/aggregation/486/486/486/4/ds02/open/xml
</cdiUrl>

Figure 10 CDI collections inventory document

The inventory document is an XML document with direct links to the individual dataset collections (ca 900 + records). Each dataset collection is described with a document compliant with the latest SeaDataNet CDI metadata standard[11].

Each SeaDataNet CDI XML document is mapped to a Blue-Cloud collection record. In the following table the mapping towards the Blue-Cloud metadata core elements is reported.

Blue-Cloud core metadata element	SeaDataNet Open CDI metadata element
Identifier	/gmd:MD_Metadata/gmd:fileIdentifier
Title	/gmd:MD_Metadata/ gmd:identificationInfo/sdn:SDN_DataIdentification/gmd:citation/gmd:CI_Citation/gmd:title
Keyword	/gmd:MD_Metadata/gmd:identificationInfo/sdn:SDN_DataIdentification /gmd:descriptiveKeywords/gmd:MD_Keywords/gmd:keyword
Bounding box	/gmd:MD_Metadata/gmd:identificationInfo/sdn:SDN_DataIdentification /gmd:extent/gmd:EX_Extent/gmd:geographicElement/gmd:EX_GeographicBoundingBox
Temporal extent	/gmd:MD_Metadata/gmd:identificationInfo/sdn:SDN_DataIdentification/gmd:extent /gmd:EX_Extent/gmd:temporalElement/gmd:EX_TemporalExtent/gmd:extent
Parameter	/gmd:MD_Metadata/gmd:identificationInfo/sdn:SDN_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords/gmd:keyword/sdn:SDN_ParameterDiscoveryCode

[11] *SeaDataNet CDI metadata standard https://www.seadatanet.org/Standards/Metadata-formats/CDI*

Instrument	/gmd:MD_Metadata/gmd:identificationInfo/sdn:SDN_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords/gmd:keyword/sdn:SDN_DeviceCategoryCode
Platform	/gmd:MD_Metadata/gmd:identificationInfo/sdn:SDN_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords/gmd:keyword/sdn:SDN_PlatformCategoryCode
Organization	/gmd:MD_Metadata/gmd:distributionInfo/gmd:MD_Distribution/gmd:distributor/gmd:MD_Distributor/gmd:distributorContact/gmd:CI_ResponsibleParty/gmd:organisationName/sdn:SDN_EDMOCode
Date stamp	/gmd:MD_Metadata/gmd:dateStamp
Revision date	/gmd:MD_Metadata/gmd:identificationInfo/sdn:SDN_DataIdentification/gmd:citation/gmd:CI_Citation/gmd:date/gmd:CI_Date[gmd:dateType/gmd:CI_DateTypeCode/@codeListValue='revision']/gmd:date

Table 4: mapping of SeaDataNet CDI metadata elements to Blue-Cloud core elements

Level 1:

As a result, the DAB broker publishes SeaDataNet CDI collections through the following endpoint:

https://blue-cloud.geodab.eu/gs-service/services/essi/view/seadatanet-open/csw

with following GetCapabilities

https://blue-cloud.geodab.eu/gs-service/services/essi/view/seadatanet-open/csw?service=CSW&request=GetCapabilities&version=2.0.2

The returned records are expressed according to the Blue-Cloud metadata profile, that is a ISO 19115 based metadata profile encoded using the recent ISO 19115-3:2016 XML schema.

To assess the quantity and quality of the overall Blue-Cloud metadata content a metadata report is made available at:

https://dabreporting.s3.amazonaws.com/BlueCloud/BlueCloudReport_brief.html

The report shows with graphical indicators the most recent status of the Blue-Cloud metadata content.

SeaDataNet Open available service at: **https://blue-cloud.geodab.eu/gs-service/services/essi/view/seadatanet-open/csw**
SeaDataNet Open available test portal at: https://blue-cloud.geodab.eu/gs-service/search?view=seadatanet-open
Total number of records: **922** Number of records analyzed: **922** Percentage of records analyzed: **100%**

Metadata element	Path	Completeness
IDENTIFIER	//gmd:fileIdentifier/gco:CharacterString	100%
TITLE	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:citation/gmd:CI_Citation/gmd:title/*[1]	100%
KEYWORD	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords[not(gmd:type) or not(contains('platform instrument',gmd:type/gmd:MD_KeywordTypeCode/@codeListValue))]/gmd:keyword/*[1]	100%
BOUNDING_BOX	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:geographicElement /gmd:EX_GeographicBoundingBox/gmd:westBoundLongitude/gco:Decimal /gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:geographicElement /gmd:EX_GeographicBoundingBox/gmd:eastBoundLongitude/gco:Decimal /gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:geographicElement /gmd:EX_GeographicBoundingBox/gmd:southBoundLatitude/gco:Decimal /gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:geographicElement /gmd:EX_GeographicBoundingBox/gmd:northBoundLatitude/gco:Decimal	100%
TEMPORAL_EXTENT	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:temporalElement/gmd:EX_TemporalExtent /gmd:extent/gml32:TimePeriod/gml32:beginPosition /gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:temporalElement/gmd:EX_TemporalExtent /gmd:extent/gml32:TimePeriod/gml32:endPosition	100%
PARAMETER	/gmi2019:MI_Metadata/gmd:contentInfo/gmi2019:MI_CoverageDescription/gmd:attributeDescription/gco:RecordType	100%
INSTRUMENT	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords[gmd:type /gmd:MD_KeywordTypeCode/@codeListValue='instrument']/gmd:keyword/*[1]	100%
PLATFORM	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords[gmd:type /gmd:MD_KeywordTypeCode/@codeListValue='platform']/gmd:keyword/*[1]	100%
ORGANIZATION	//gmd:CI_ResponsibleParty/gmd:organisationName/*[1]	100%
DATESTAMP	/gmi2019:MI_Metadata/gmd:dateStamp/gco:Date	100%
REVISION_DATE	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:citation/gmd:CI_Citation/gmd:date/gmd:CI_Date[gmd:dateType /gmd:CI_DateTypeCode/@codeListValue='revision']/gmd:date/gco:Date	100%
RESOURCE_IDENTIFIER	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:citation/gmd:CI_Citation/gmd:identifier/gmd:MD_Identifier/gmd:code /gco:CharacterString	100%

Figure 11 DAB completeness report of core metadata elements in the SeaDataNet CDI open collections service as determined from their web service

The output of the Blue-Cloud DAB broker is directly used to drive the level 1 discovery service of the Blue-Cloud Data Discovery & Access service.

Level 2:

At granular level there is a dedicated API for the SeaDataNet CDI service, in particular for open data, and this works in combination with the CDI shopping mechanism. That SeaDataNet shopping mechanism works with Marine-ID as AAI service, which is also used for the Blue-Cloud DD&AS shopping mechanism. The connection between the DD&AS and CDI API is arranged by having a Marine-ID for the Blue-Cloud as a CDI user. The CDI open API is implemented as a Swagger-API, which includes documentation and options to try out and build queries step by step. The dialogue itself is similar to the human User Interface of the CDI service.

https://cdi-open.seadatanet.org/api

The API supports the model of direct facets and queries, without having to build a new interface on harvested metadata. This CDI API facilitates in Blue-Cloud granular searching by:

- Free search
- Lat-Lon box
- Date period
- Measuring area type
- Discovery Parameter (P02)
- Parameter Group (P03)
- Discipline (P08)
- Point of Contact (EDMO)
- Point of Contact Country (ISO3166)
- Data Originator (EDMO)
- Data Custodian (EDMO)
- Data Distributor (EDMO)

The data links are direct shopping requests which are handled by the SeaDataNet CDI service and then as data packages transferred to the Blue-Cloud data delivery service.

4.3.3 Proposed optimisation

The SeaDataNet CDI Data Discovery & Access service is a fully operational service which is used daily by data providers and data users, also because it is one of the major pillars under EMODnet. The SeaDataNet CDI service is one of the well established services of the SeaDataNet AISBL pan-European network of NODCs. It has more than 110 data centre nodes, of which circa 60 are very active with updating and submitting new metadata and data. The SeaDataNet CDI service, together with the other SeaDataNet pan-European directories and vocabularies, is used as the data management infrastructure for several EMODnet thematic groups, such as EMODnet Chemistry (focus on eutrophication, contaminants, and marine litter data), EMODnet Bathymetry (focus on bathymetry survey data sets), and EMODnet Physics (focus on physical data sets). Moreover, it is a major data management infrastructure for EMODnet Ingestion, as most submitted data sets after elaboration to standard formats are ingested into the SeaDataNet CDI service.

The CDI service has online components for import of new and updated metadata and data submissions from connected data providers and for export by means of discovery and access through a shopping basket mechanism for registered users. These components are operated, managed and maintained by MARIS together with EUDAT service providers. The CDI service also has offline components, consisting of dedicated software packages for editing new CDI entries (MIKADO), converting data sets to SeaDataNet formats (NEMO), checking data formats (OCTOPUS), and connecting data centres to the online import service part for ingesting new and updated metadata and data (REPLICATION MANAGER). These offline

components are developed, managed, and maintained by Ifremer. There is regular maintenance of most components, following feedback from data providers and users, and following FAIRness improvements, new data types, practical issues, etc.

For marking up and enriching the metadata XML and data sets, there is machine-to-machine interconnection with the SeaDataNet pan-European directories and controlled vocabularies, using linked data principles. These directories are operated and managed by MARIS (EDMO, EDMERP, EDIOS), BODC (EDMED, EDIOS, Vocabularies), and Ifremer (CSR). Their population is done by the network of NODCs, while the Vocabularies have a wider global governance scheme as the SeaDataNet controlled vocabularies are used on a global scale by many. The SeaDataNet directories and vocabularies are also increasingly used in the European operational oceanography community (Copernicus Marine INSTAC, EuroGOOS).

Overall, the SeaDataNet CDI service is well established and embedded in the European marine community as one of the leading infrastructures for marine and ocean data management. As such it has long term perspective, while it is sustained for coming years as part of EMODnet, DTO, and other European top down initiatives. At the same time, more efforts and associated funding will be needed in the coming years to enlarge the workforce of data managers at the NODCs and regularly innovate the core systems used. This will be necessary to be able to deal with the 'flood' of new data, that is being acquired and resulting from many new EU research projects, citizen science initiatives, smart and low cost sensors, increased monitoring, engaging more groups in society for sharing their data etc. Moreover, EU strategy is aimed at mobilising more FAIR data sharing and directing data providers to established European infrastructures such as SeaDataNet in order to have access to a far richer base of data and enabling new challenges such as foreseen with Destination-E and Digital Twins of the Ocean (DTO).

As part of the Blue-Cloud 2026 optimisation can and will be undertaken on the following aspects:

- Re-organisation and upgrading of the cloud components of the online service
- Semantic interoperability

The latest SeaDataNet CDI service, version 5, makes use of a cloud infrastructure at EUDAT. This infrastructure has been designed and deployed during the SeaDataCloud project (2016-2021), and was launched in October 2019. The kernel is a cloud database, in which all unrestricted data sets (CC-BY-4.0) are replicated, and which is used to serve user requests from a central database, without having to go back to each of the individual data centres as was the case in earlier versions of the SeaDataNet CDI service. The cloud database is populated through the online import service and data requests are handled through the export service, which is driven by the shopping mechanism. The central CDI meta database, import and export services, discovery and access (shopping) services, and the overall master system are

hosted and managed by MARIS, while the cloud system is operated by EUDAT partners. In current practice, CINECA hosts and manages the development environment, while CSC hosts the test and production environments, including the central data cloud. Over time, a number of issues have appeared, such as in particular:

- for the cloud database use is made of B2SAFE - iRODS, as that supports synchronisation between cloud databases ; in operational practice, there are disturbances caused by iRODS which give 'hanging' transactions and required rebooting. For that reason, EUDAT is phasing out the use of iRODS in its various systems ;
- importing of new or upgraded data sets into the cloud system is done in batches and involves using B2HANDLE for assigning Persistent Identifiers (PIDs) for each data set in a batch so that it can be used as coupling element between the CDI metadata and the imported data files. The PID assigning process runs at another EUDAT partner (SURF) and in operational practice, it takes relatively a lot of time when a data provider submits a batch of thousands of new data sets which happens regularly. The PID assigning process then slows down the import of multiple batches considerably.

As part of the Blue-Cloud 2026 it has been decided with EUDAT to re-organise the set-up of the SeaDataNet CDI cloud infrastructure by concentrating all kernel services and all environments (development; test; production) at CINECA and no longer partly at CSC. Moreover, CINECA has designed a new set-up for those kernel services, making optimal use of its recently deployed new cloud infrastructure.

A plan has been made by CINECA which respects all the existing APIs which support the communication exchanges between the CDI service components hosted and managed by MARIS, while the architecture and implementation for the internal cloud services have been re-designed.

Part of the plan is to replace the iRODS system with a normal filesystem mounted as a volume in a Virtual Machine on the CINECA cloud. The filesystem will be divided into three different main directories to mimic the three IRODS' main collections : one related to the batches to ingest, one related to the data in productions, and one related to the orders. One directory will be used for storing the batches before they are moved to production, one will host the files that are already in production and the third one will store the different orders.

The B2HANDLE – PID system will be replaced by a system of local Unique identifiers (UID) that will be assigned to the single resources in production. These UIDs will maintain the same prefixes previously used to identify the type of resources, in order to mimic the previous rules used for PID assignment. The UIDs

of the resources already existing at the moment of the migration to the new system will match the PIDs they had in the old system to not create inconsistency between the old system and the new one. The correspondence between the UID and the path of the file will be saved in a local database.

In the meantime, CINECA has developed the new system components for deploying the services, which are currently in a new development environment at the CINECA cloud infrastructure.

There are regular meetings between MARIS, CINECA, and CSC to discuss and monitor the further development, testing, and deployments, aimed at fully migrating from CSC to CINECA and switching to the new CDI service Version 6 by end 2023.

Another aspect in the optimisation in Blue-Cloud 2026 is semantic interoperability. The SeaDataNet CDI format is a dedicated marine profile upon the ISO19115 – 19139 metadata standard and makes already optimal use of SeaDataNet controlled vocabularies and European directories (CSR, EDMED, EDMERP, EDMO) for marking up most metadata tags. For that purpose, the SeaDataNet CDI schema has been expanded with Schematron to support inclusion of declarations for semantic terms and their vocabulary sources, which are used when importing and parsing CDI XML records. The CDI output for both the CDI aggregation level and the granular level in XML carries these semantic declarations, so that receiving services could make use of this for semantic interoperabilty.

The following gives an example of a CDI record with vocabularies, EDMO, EDMERP, EDMED, and CSR relations:

https://cdi.seadatanet.org/report/1508063

its XML output, which contains all vocab declarations:

https://cdi.seadatanet.org/report/1508063/xml

while the json output gives :

https://cdi.seadatanet.org/report/1508063/json

This is incomplete and it is proposed to add also a JSON-LD output which will include all vocabulary declarations. The following example from MEDIN might be adopted :

```
"keywords": [
  {
    "@type": "DefinedTerm",
    "inDefinedTermSet": "https://vocab.nerc.ac.uk/collection/L13/current/",
```

"termCode": " http://vocab.nerc.ac.uk/collection/L13/current/SD”
"name": "sediment"
},
]

Figure 12 Possible triple in JSON-LD to include literal term but also its vocabulary term and the associated vocabulary service

4.3.4 Planned actions

ACTION : Make further steps with the upgrading of the cloud system part of the SeaDataNet CDI service, migrating from CSC to CINECA. Steps will include testing the new developed services, starting with import and export, PID assignments, and use of QA-QC docker containers. In parallel, CINECA will set up the test and production environments next to the development environment. In addition, there will be testing of migration and ingestion of existing data sets from the CSC cloud to CINECA cloud, and finding a solution for moving outstanding orders from CSC to CINECA without disrupting services. Another step is arranging and testing the interaction between Marine-ID AAI and the AAI system as internally used by CINECA. Finally, solutions have to be deployed for reliable back-up of the cloud data and related metadata, and for monitoring the operational functioning of the new central cloud components. These actions will be undertaken by CINECA, MARIS, and CSC and aiming for release by end 2023, which might be challenging.

ACTION : MARIS together with CNR-IIA to formulate and deploy a CDI JSON-LD output including declarations of literal, coding and associated vocabularies as used at the SeaDataNet CDI metadata format. Apply solutions as found by CNR-IIA for including and propagating these extra vocab declarations in the common Blue-Cloud DAB metadata format and DAB broker service. The latter action is combined with the earlier actions for EMODnet Chemistry products and SeaDataNet data products.

4.4. EurOBIS – EMODnet Biology data collections service

The EMODnet Biology is another thematic network within EMODnet. It aims at providing open and free access to interoperable data and data products on temporal and spatial distribution of marine species (angiosperms, benthos, birds, fish, macroalgae, mammals, phytoplankton, reptiles, zooplankton) and species traits from European regional seas, as defined by the EEA's 'Europe's seas' dataset (Arctic Ocean, (North) Atlantic Ocean, Baltic Sea, Black Sea, Mediterranean Sea and North Sea). EMODnet Biology's taxonomic backbone is built upon the World Register of Marine Species (WoRMS[12]) and supported by the European Ocean Biodiversity Information System (EurOBIS[13]) data infrastructure, with tools and services developed in collaboration with Lifewatch ERIC and Lifewatch Marine. The federation in Blue-Cloud Data Discovery & Access Service concerns the EurOBIS service.

4.4.1 Description

EurOBIS was developed by the Flanders Marine Institute (VLIZ) 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 1300+ datasets, representing > 62.000 species and circa 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, redating Linnaeus and Darwin. 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. Similar to what happens with other regional nodes, EurOBIS data flow to the global initiative Ocean Biogeographic nformation System (OBIS[14]) and eventually become available via the Global Biodiversity Information Facility (GBIF[15]), hosting global marine and terrestrial distribution data.

[12] https://www.marinespecies.org

[13] https://www.eurobis.org

[14] https://obis.org

[15] https://www.gbif.org

4.4.2 Current federation

Initially, use was made of the DCAT service of EurOBIS which gives Ecological Metadata Language (EML) output:

https://ipt.vliz.be/eurobis/dcat

From the obtained dataset URLs, individual EML documents can be obtained by executing requests such as the following:

https://ipt.vliz.be/eurobis/eml?r=idod_ipms_phaeo#Dataset

Each EML document is then mapped to a Blue-Cloud collection record and becomes part of the Blue-Cloud GeoDAB CSW service, that is used to drive the level 1 of the Blue-Cloud Data Discovery & Access service. Each data set has a unique and persistent 'dasid'. This dasid is used to retrieve data sets. EurOBIS users are interested in only 1 level, the collections, which each can contain millions of observations. The direct download URL has been checked to work for all the datasets and is always returning a zipped package.

As part of the Blue-Cloud 2026 discussions with VLIZ, it was decided to migrate from the **DCAT service** to a newly established **OAI-PMH service** of EurOBIS.

https://www.eurobis.org/oai/?verb=Identify

This also gives records in EML format. An EML example:

https://www.eurobis.org/oai/?verb=GetRecord&identifier=oai:marineinfo.org:id:dataset:5951&metadataPrefix=eml

This service is completely covering all published EurOBIS collection records, while the DCAT was missing a number of records. The new output also contains the ‘dasid’ which is needed for retrieving the data collections sets.

In the meantime, the OAI-PMH service has been successfully harvested by the Blue-Cloud DAB service of CNR and results are available in the DAB OGC CSW.

In the following table the mapping towards the Blue-Cloud metadata core elements is reported.

Blue-Cloud core metadata element	EurOBIS metadata element
Identifier	GetRecord/record/header/identifier
Title	/eml/dataset/title

Blue-Cloud core metadata element	EurOBIS metadata element
Keyword	/keywordSet/keyword
Bounding box	/geographicCoverage/geographicDescription/boundingCoordinates/westBoundingCoordinate
	/geographicCoverage/geographicDescription/boundingCoordinates/eastBoundingCoordinate
	/geographicCoverage/geographicDescription/boundingCoordinates/southBoundingCoordinate
	/geographicCoverage/geographicDescription/boundingCoordinates/northBoundingCoordinate
Temporal extent	/temporalCoverage/rangeOfDates/beginDate/calendarDate
	/temporalCoverage/rangeOfDates/endDate/calendarDate
Parameter	/additionalMetadata/describes/para/describes/metadata/parameter /taxonomicCoverage/generalTaxonomicCoverage/generalTaxonomicCoverage/taxonomicClassification/taxonRankName/taxonRankValue
Instrument	/methods/methodStep/description/title/para/instrumentation
Platform	/additionalMetadata/describes/para/describes/metadata/platform
Organization	/eml/dataset/creator/organizationName /eml/dataset/metadataProvider/organizationName /eml/dataset/associatedParty/organizationName /eml/dataset/contact/organizationName
Date stamp	/additionalMetadata/metadata/gbif/dateStamp
	Temporal extent end
Revision date	Temporal extent end

Table 5: mapping of EurOBIS metadata elements to Blue-Cloud core elements

Notes:

- for bounding box, only the "EurOBIS calculated BBOX" is used for retrieving the bounding box
- there are hardly any EurOBIS records with instrumentation

Level 1:

As a result, the DAB broker publishes EurOBIS collection records through the following endpoint:

https://blue-cloud.geodab.eu/gs-service/services/essi/view/eurobis/csw

with following GetCapabilities

https://blue-cloud.geodab.eu/gs-service/services/essi/view/eurobis/csw?service=CSW&request=GetCapabilities&version=2.0.2

The DAB Report on completeness can be found at:

https://dabreporting.s3.amazonaws.com/BlueCloud/BlueCloudReport_full.html

EurOBIS available service at: **https://blue-cloud.geodab.eu/gs-service/services/essi/view/eurobis/csw**
EurOBIS available test portal at: https://blue-cloud.geodab.eu/gs-service/search?view=eurobis
Total number of records: **1302** Number of records analyzed: **1302** Percentage of records analyzed: **100%**

Metadata element	Path	Completeness
IDENTIFIER	//gmd:fileIdentifier/gco:CharacterString	100%
TITLE	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:citation/gmd:CI_Citation/gmd:title/*[1]	100%
KEYWORD	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords[not(gmd:type) or not(contains('platform instrument',gmd:type/gmd:MD_KeywordTypeCode/@codeListValue))]/gmd:keyword/*[1]	100%
BOUNDING_BOX	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:geographicElement /gmd:EX_GeographicBoundingBox/gmd:westBoundLongitude/gco:Decimal /gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:geographicElement /gmd:EX_GeographicBoundingBox/gmd:eastBoundLongitude/gco:Decimal /gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:geographicElement /gmd:EX_GeographicBoundingBox/gmd:southBoundLatitude/gco:Decimal /gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:geographicElement /gmd:EX_GeographicBoundingBox/gmd:northBoundLatitude/gco:Decimal	84%
TEMPORAL_EXTENT	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:temporalElement/gmd:EX_TemporalExtent /gmd:extent/gml32:TimePeriod/gml32:beginPosition /gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:temporalElement/gmd:EX_TemporalExtent /gmd:extent/gml32:TimePeriod/gml32:endPosition	91%
PARAMETER	/gmi2019:MI_Metadata/gmd:contentInfo/gmi2019:MI_CoverageDescription/gmd:attributeDescription/gco:RecordType	86%
INSTRUMENT	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords[gmd:type /gmd:MD_KeywordTypeCode/@codeListValue='instrument']/gmd:keyword/*[1]	0%
PLATFORM	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords[gmd:type /gmd:MD_KeywordTypeCode/@codeListValue='platform']/gmd:keyword/*[1]	19%
ORGANIZATION	//gmd:CI_ResponsibleParty/gmd:organisationName/*[1]	100%
DATESTAMP	/gmi2019:MI_Metadata/gmd:dateStamp/gco:Date	100%
REVISION_DATE	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:citation/gmd:CI_Citation/gmd:date/gmd:CI_Date[gmd:dateType /gmd:CI_DateTypeCode/@codeListValue='revision']/gmd:date/gco:Date	100%

Figure 13 DAB completeness report of core metadata elements in the EurOBIS OAI-PMH service as determined from their web service

The output of the Blue-Cloud DAB broker is directly used to drive the level 1 discovery service of the Blue-Cloud Data Discovery & Access service.

Level 2:

In this case, there is only a collections level and therefore the level search and actual downloading is also derived from the Blue-Cloud DAB broker service output. For the level 2 search extra search criteria from the common profile are used, next to the universal criteria of level 1:

- Free search
- Lat-Lon box
- Date period
- Keywords
- Parameters
- Organisation

The download links can be found as GeoServer WFS URL. For instance:

http://geo.vliz.be/geoserver/wfs/ows?service=WFS&version=1.1.0&request=GetFeature&typeName=Dataportal:eurobis-obisenv_basic&resultType=results&viewParams=where:datasetid+IN+(1002);context:0100&outputFormat=csv

4.4.3 Proposed optimisation

The EurOBIS service is a fully operational service which is used daily by data providers and data users, also because it is one of the major pillars under EMODnet and functioning as European node in the global OBIS and GBIF infrastructures. As part of EMODnet Biology there is guaranteed funding for maintaining and updating new records to EurOBIS in the coming years and perspectives for the long term. The EurOBIS service is also fully guaranteed by its main operator and manager VLIZ as it represents a core and valued service for VLIZ together with WoRMS and MarineRegions in the European marine research community, and in particular context with EMODnet and LifeWatch.

Optimisation is considered in a number of aspects :

- Further checking and completing of existing EurOBIS records for all fields in the Blue-Cloud common metadata profile
- Optimising the harvesting process for updating the DAB broker catalogue and CSW
- Improving semantic interoperability

Currently, CNR-IIA runs a weekly harvesting of the OAI-PMH endpoint for a full replacement of the DAB catalogue for EurOBIS. At present, this contains circa 1300+ records, which is not a big burden for harvesting. It will be analysed if selective harvesting for only new and/or updated records is feasible, whereby also the case of de-activated records must be taken into account. At present, the harvesting report is available online. It might be useful to distribute such report also by email after each email to the repository managers, which will make them more informed and aware of content deficiencies that could be corrected.

For improving the semantic interoperability, there are the following considerations. The current method used for Blue-Cloud is harvesting EML records from the EurOBIS OAI-PMH endpoint. However, VLIZ is managing and maintaining the IMIS system as marineinfo.org metadata system. This system is feeding into EurOBIS. IMIS is more flexible and VLIZ is underway with developments for improving export formats. A templating approach is adopted to output the datasets metadata from the IMIS database to RDF in various serialisations, with JSON-LD and Turtle as options. JSON-LD is more flexible than EML and allows more easily to include semantic triples with literal terms, code of terms, and associated vocabularies for improving the semantic interoperability. The JSON-LD could be provided with a REST service, but the JSON-LD could also be converted to XML, which then could be served through the OAI-PMH protocol service. Both options, JSON-LD or XML, are feasible for the DAB brokerage service to harvest richer metadata including full semantics.

These developments are planned for the VLIZ IMIS system in the framework of the FAIR-EASE project and can be adopted to optimize the EurOBIS harvesting.

4.4.4 Planned actions

ACTION : To work on getting more complete contents for EurOBIS and updating the metadata for missing/incomplete fields (parameters, platforms, instruments), by pulling them from the EurOBIS database into the marineinfo.org metadata system (IMIS) and to further update these metadata for missing fields that cannot be automatically harvested, but need to be obtained manually (e.g. from data providers, publications, etc). To be undertaken by VLIZ and EurOBIS team.

ACTION : To develop and provide JSON-LD metadata export from EurOBIS by REST service or alternatively as reformatted into XML via OAI-PMH endpoint. IN parallel, further improving the semantics of the metadata in the marineinfo.org metadata system (IMIS), in particular for fields relevant for the Blue-Cloud federation. To be undertaken by VLIZ.

ACTION: To explore if selective harvesting of EurOBIS for new, updated, and de-activated records is feasible, and additional distribution of harvesting reports by email. To be undertaken by CNR-IIA.

4.5. EuroArgo – Argo GDAC data service

The Euro-Argo 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.

4.5.1 Description

The Euro-Argo ERIC also provides access to quality-controlled data and derived products:

- **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). 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.

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 (Argos) to 1000 (Iridium) data points (vertical resolution). In total, there are currently 16.000 Argo floats which generated more than 2 million files. Metadata and data for profiles and trajectories, including technical information are made available as NetCDF (CF) files. In addition, Argo products are generated and made available as gridded fields.

4.5.2 Current federation

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.
For Blue-Cloud use is made of the following Swagger 2.0 API service which is JSON based:

https://fleetmonitoring.euro-argo.eu/

The JSON based API allows to discover metadata about the ARGO floats. For adopting the Blue-Cloud DD&AS approach, the Argo platforms are considered as collections, while the related data sets are considered as granular entries. Currently, there are 17.000+ platform entries including all historic Argo floats, while the number of related data sets is more than 2.8 Million.

ARGO "float" level records are then mapped to the Blue-Cloud first level (collection) metadata records. It is possible to harvest metadata for all the floats by issuing the following HTTP-GET request to retrieve all the float identifiers:

https://fleetmonitoring.euro-argo.eu/platformCodes

The list of available identifiers coming from the previous request are used to retrieve full metadata for each float. Example given to obtain the metadata record for float with identifier *6903238* the following HTTP-GET request is executed:

https://fleetmonitoring.euro-argo.eu/floats/6903238

In the following table the mapping towards the Blue-Cloud metadata core elements is reported for the float metadata.

Blue-Cloud core metadata element	**ARGO metadata element**
Identifier	platform_code
Title	platform_name + platform_code + platform_description
Keyword	projectName
	countryCode
	model
	maker
	deployment_cruiseName
	platform_name

Blue-Cloud core metadata element	ARGO metadata element
	sensors_model
Bounding box	cycles_lat
	cycles_lon
Temporal extent	earliestCycle_startDate
	latestCycle_startDate
Parameter	variables
Instrument	sensors_id
	sensors_model
	sensors_maker
	sensors_serial
Platform	platform_name
	platform_code
	platform_description
Organization	deployment_ principalInvestigatorName
	owner
	dataCenter_name
	institution_name
Date stamp	latestCycle_startDate
Revision date	latestCycle_startDate

Table 6: mapping of Argo metadata elements to Blue-Cloud core elements

Level 1:

As a result, the DAB broker publishes Euro-Argo / Argo collection records through the following endpoint:

https://blue-cloud.geodab.eu/gs-service/services/essi/view/argo/csw

with following GetCapabilities

https://blue-cloud.geodab.eu/gs-service/services/essi/view/argo/csw?service=CSW&request=GetCapabilities&version=2.0.2

The DAB Report on completeness can be found at:

https://dabreporting.s3.amazonaws.com/BlueCloud/BlueCloudReport_full.html

ARGO available service at: **https://blue-cloud.geodab.eu/gs-service/services/essi/view/argo/csw**
ARGO available test portal at: https://blue-cloud.geodab.eu/gs-service/search?view=argo
Total number of records: **18323** Number of records analyzed: **18323** Percentage of records analyzed: **100%**

Metadata element	Path	Completeness
IDENTIFIER	//gmd:fileIdentifier/gco:CharacterString	100%
TITLE	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:citation/gmd:CI_Citation/gmd:title/*[1]	100%
KEYWORD	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords[not(gmd:type) or not(contains('platform instrument',gmd:type/gmd:MD_KeywordTypeCode/@codeListValue))]/gmd:keyword/*[1]	100%
BOUNDING_BOX	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:geographicElement /gmd:EX_GeographicBoundingBox/gmd:westBoundLongitude/gco:Decimal /gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:geographicElement /gmd:EX_GeographicBoundingBox/gmd:eastBoundLongitude/gco:Decimal /gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:geographicElement /gmd:EX_GeographicBoundingBox/gmd:southBoundLatitude/gco:Decimal /gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:geographicElement /gmd:EX_GeographicBoundingBox/gmd:northBoundLatitude/gco:Decimal	97%
TEMPORAL_EXTENT	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:temporalElement/gmd:EX_TemporalExtent /gmd:extent/gml32:TimePeriod/gml32:beginPosition /gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:temporalElement/gmd:EX_TemporalExtent /gmd:extent/gml32:TimePeriod/gml32:endPosition	97%
PARAMETER	/gmi2019:MI_Metadata/gmd:contentInfo/gmi2019:MI_CoverageDescription/gmd:attributeDescription/gco:RecordType	99%
INSTRUMENT	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords[gmd:type /gmd:MD_KeywordTypeCode/@codeListValue='instrument']/gmd:keyword/*[1]	98%
PLATFORM	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords[gmd:type /gmd:MD_KeywordTypeCode/@codeListValue='platform']/gmd:keyword/*[1]	100%
ORGANIZATION	//gmd:CI_ResponsibleParty/gmd:organisationName/*[1]	100%
DATESTAMP	/gmi2019:MI_Metadata/gmd:dateStamp/gco:Date	97%
REVISION_DATE	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:citation/gmd:CI_Citation/gmd:date/gmd:CI_Date[gmd:dateType /gmd:CI_DateTypeCode/@codeListValue='revision']/gmd:date/gco:Date	97%
RESOURCE_IDENTIFIER	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:citation/gmd:CI_Citation/gmd:identifier/gmd:MD_Identifier/gmd:code /gco:CharacterString	100%

Figure 14 DAB completeness report of core metadata elements in the Euro-Argo dashboard as determined from their web service

The output of the Blue-Cloud DAB broker is directly used to drive the level 1 discovery service of the Blue-Cloud Data Discovery & Access service. A filter is applied to exclude platform codes starting with "EXUD" as these will have no data sets.

Level 2:

For level 2, a harvest is made by MARIS of the full JSON information for all Argo floats as indexed at Level 1. This is done by harvesting the info for each Argo platform from the API:

https://fleetmonitoring.euro-argo.eu/platformCodes

This metadata is stored in an elastic index because the fleetmonitoring API doesn't support searching the way DD&AS needs. To get the data for a float cycle, the following is done:

E.g. for AOML, platform 13859, cycle 127; request ftp directory listing with:

ftp://ftp.ifremer.fr/ifremer/argo/dac/AOML/13859/profiles/*_127*

(notice the *_127* wildcard filename)

This will give us all the datafiles for cycle 127, e.g.:

ftp://ftp.ifremer.fr/ifremer/argo/dac/aoml/13859/profiles/R13859_127.nc

For level 2 the following search criteria are provided from the Elastic database:

- Free search
- Lat-Lon box
- Date period
- Platform type
- Country
- Status code
- Year of deployment
- Transmission system
- Data centre name
- Variables
- Nerworks
- Owners
- Projects

Most criteria are provided as search facets, next to the free search, date search and geographic search. The details per float and cycle provide additional details.

4.5.3 Proposed optimisation

The Euro-Argo-Argo service is a fully operational service which is used daily by data providers and data users. It is a major global ocean observing system and there are many use applications for the Argo data sets. Moreover, Argo is an international initiative with Euro-Argo as European contribution, supported through the ERIC legal structure by several European countries with funding at government level and operation at research institutes. This implicates that the observation network and Euro-Argo have a sustained long term operation, although there is still need for additional floats for better and sustained coverage of the oceans and seas and for expanding the fleet with new floats that carry multiple instruments and collect more types of parameters, in particular for BioGeoChemistry next to classic Physics.

The API as made available at the dashboard is an important service for Euro-Argo to provide discovery and access and therefore will also be sustained and further improved, where needed and possible.

Optimisation is considered in a number of aspects :

- Ensuring that the platforms as used at level 2 are synchronised with the list of platforms that result from the harvesting through the DAB service and the filter validation.
- Improving semantic interoperabiity

Euro-Argo makes use of a number of vocabularies for harmonising metadata and data descriptions. These vocabularies[16] are hosted and part of the Nerc Vocabulary Services (NVS) which is also hosting the SeaDataNet controlled vocabularies.

However, in the JSON output from the API, only the literal descriptions are given. It should be easy to expand the JSON output with the triples giving literal description, term code, and associated vocabulary url. This enriched output then could be harvested and included in the Blue-Cloud DAB catalogue and OGC CSW service as well as used at level 2.

4.5.4 Planned actions

ACTION : To synchronise the list of Argo platforms at level 2 with the list at level 1 after filtering. To be done by MARIS on short term.

ACTION : To expand the JSON output from the Argo fleetmonitoring API with triples for lieral description, coding of terms, and associated vocabulary URLs for optimising semantic interoperability. To be undertaken by Ifremer.

[16] *https://www.argodatamgt.org/Documentation/Argo-vocabulary-server*

4.6. ELIXIR – ENA data service

4.6.1 Description

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.

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.

4.6.2 Current federation

A Swagger 2.0 based API is published by EMBL-EBI to discover metadata about the available studies. The output of the API is JSON:

https://www.ebi.ac.uk/ena/portal/api/

ELIXIR-ENA "study" level records are mapped to the Blue-Cloud first level (collection) metadata records. For this, a set of predefined studies and study collections are selected to be harvested. These include the following ones:

- • Tara Oceans Metagenome (PRJEB402)
- • Ocean Sampling Day (PRJEB5129)
- • Malaspina (PRJNA330770)

For each study, a metadata record is retrieved describing it, using the query operation with the specified study identifier and selecting "study" as the result type. The returned study-level metadata elements are often lacking some important metadata elements (most notably the bounding box), so to augment them, also sample-level metadata documents are retrieved (which contain the latitude and longitude of the specific acquisition), to complement study-level metadata. Again, the query operation is used, but this time selecting "read_run" as the result type

In the following table the mapping towards the Blue-Cloud metadata core elements is reported.

Funded by the European Union
Blue-Cloud 2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227

Blue-Cloud core metadata element	ELIXIR-ENA metadata element
Identifier	study_accession
Title	study_title
Keyword	keywords
	environment_biome (from sample records)
	environment_feature (from sample records)
	environment_material (from sample records)
	environmental_package (from sample records)
	investigation_type (from sample records)
	country (from sample records)
	sample_alias (from sample records)
	project_name (from sample records)
Bounding Box	lat (from sample records)
	lon (from sample records)
	location (from sample records)
Temporal extent	last_updated
	last_updated (from sample records)
	first_created (from sample records)
	collection_date (from sample records)
Parameter	scientific_name (from sample records)
Instrument	instrument_model (from sample records)
	sequencing_method (from sample records)
Platform	sampling_platform (from sample records)
	instrument_platform (from sample records)
Organization	center_name
Date stamp	last_updated
	temporal extent end position
Revision date	same as date stamp

Table 7: mapping of ELIXIR ENA metadata elements to Blue-Cloud core elements

Level 1:

As a result, the DAB broker publishes ELIXIR-ENA collection records through the following endpoint:

https://blue-cloud.geodab.eu/gs-service/services/essi/view/elixir-ena/csw

with following GetCapabilities

https://blue-cloud.geodab.eu/gs-service/services/essi/view/elixir-ena/csw?service=CSW&request=GetCapabilities&version=2.0.2

The DAB Report on completeness can be found at:

https://dabreporting.s3.amazonaws.com/BlueCloud/BlueCloudReport_full.html

ELIXIR-ENA available service at: **https://blue-cloud.geodab.eu/gs-service/services/essi/view/elixir-ena/csw**
ELIXIR-ENA available test portal at: https://blue-cloud.geodab.eu/gs-service/search?view=elixir-ena
Total number of records: **32** Number of records analyzed: **32** Percentage of records analyzed: **100%**

Metadata element	Path	Completeness
IDENTIFIER	//gmd:fileIdentifier/gco:CharacterString	100%
TITLE	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:citation/gmd:CI_Citation/gmd:title/*[1]	100%
KEYWORD	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords[not(gmd:type) or not(contains('platform instrument',gmd:type/gmd:MD_KeywordTypeCode/@codeListValue))]/gmd:keyword/*[1]	25%
BOUNDING_BOX	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:geographicElement /gmd:EX_GeographicBoundingBox/gmd:westBoundLongitude/gco:Decimal /gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:geographicElement /gmd:EX_GeographicBoundingBox/gmd:eastBoundLongitude/gco:Decimal /gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:geographicElement /gmd:EX_GeographicBoundingBox/gmd:southBoundLatitude/gco:Decimal /gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:geographicElement /gmd:EX_GeographicBoundingBox/gmd:northBoundLatitude/gco:Decimal	65%
TEMPORAL_EXTENT	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:temporalElement/gmd:EX_TemporalExtent /gmd:extent/gml32:TimePeriod/gml32:beginPosition /gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:temporalElement/gmd:EX_TemporalExtent /gmd:extent/gml32:TimePeriod/gml32:endPosition	75%
PARAMETER	/gmi2019:MI_Metadata/gmd:contentInfo/gmi2019:MI_CoverageDescription/gmd:attributeDescription/gco:RecordType	75%
INSTRUMENT	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords[gmd:type /gmd:MD_KeywordTypeCode/@codeListValue='instrument']/gmd:keyword/*[1]	75%
PLATFORM	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords[gmd:type /gmd:MD_KeywordTypeCode/@codeListValue='platform']/gmd:keyword/*[1]	75%
ORGANIZATION	//gmd:CI_ResponsibleParty/gmd:organisationName/*[1]	100%
DATESTAMP	/gmi2019:MI_Metadata/gmd:dateStamp/gco:Date	100%
REVISION_DATE	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:citation/gmd:CI_Citation/gmd:date/gmd:CI_Date[gmd:dateType /gmd:CI_DateTypeCode/@codeListValue='revision']/gmd:date/gco:Date	100%
RESOURCE_IDENTIFIER	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:citation/gmd:CI_Citation/gmd:identifier/gmd:MD_Identifier/gmd:code /gco:CharacterString	100%

Figure 15 DAB completeness report of core metadata elements in the ELIXIR-ENA service as determined from their web service

The current output of the Blue-Cloud DAB broker, namely 32 records related to the 3 main projects, is directly used to drive the level 1 discovery service of the Blue-Cloud Data Discovery & Access service.

Level 2:

For level 2, MARIS harvests API JSON results at interval and includes the JSON information in an Elastic index (cache). In total, information from 32 marine Projects is harvested.

The reason for harvesting/caching is, that the API was too slow and inconsistent at times during COVID, when the ENA system was largely in demand by many other users.

https://www.ebi.ac.uk/ena/portal/api/search

Funded by the European Union
Blue-Cloud 2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227

A sample query to get a data link is:

https://www.ebi.ac.uk/ena/portal/api/filereport?accession=SAMEA2623868&result=read_run&format=json

The following query was used to get the level 2 records:

```
curl --location 'https://www.ebi.ac.uk/ena/portal/api/search' \

      --header 'Content-Type: application/x-www-form-urlencoded' \

      --data-urlencode 'fields=all' \

      --data-urlencode 'result=read_study' \

      --data-urlencode 'format=json' \

    --data-urlencode 'query=(tax_tree(408172) OR environment_biome="*marine*"
OR   isolation_source="*marine*")   AND   ((study_accession="PRJEB1787")   OR
(study_accession="PRJEB1788")     OR     (study_accession="PRJEB36282")     OR
(study_accession="PRJEB36283")    OR     (study_accession="PRJEB36284")     OR
(study_accession="PRJEB36285")    OR     (study_accession="PRJEB4352")      OR
(study_accession="PRJEB4357")     OR     (study_accession="PRJEB4419")      OR
(study_accession="PRJEB4422")     OR     (study_accession="PRJEB6603")      OR
(study_accession="PRJEB6604")     OR     (study_accession="PRJEB6605")      OR
(study_accession="PRJEB6606")     OR     (study_accession="PRJEB6607")      OR
(study_accession="PRJEB6608")     OR     (study_accession="PRJEB6609")      OR
(study_accession="PRJEB6610")     OR     (study_accession="PRJEB7315")      OR
(study_accession="PRJEB7988")     OR     (study_accession="PRJEB8682")      OR
(study_accession="PRJEB9691")     OR     (study_accession="PRJEB9694")      OR
(study_accession="PRJEB9737")     OR     (study_accession="PRJEB9738")      OR
(study_accession="PRJEB9739")     OR     (study_accession="PRJEB9740")      OR
(study_accession="PRJEB9741")     OR     (study_accession="PRJEB9742")      OR
(study_accession="PRJNA326480")   OR     (study_accession="PRJNA330770")    OR
(study_accession="PRJNA477650"))'
```

This will retrieve about 6800+ records.

For level 2 the following search criteria are provided from the Elastic database:

- Free search
- Lat-Lon box

- Date period
- Study accession
- Study title
- Sampling Campaign
- Sampling Platform
- Project Name

Most criteria are provided as search facets, next to the free search, date search and geographic search. Currently, level 2 results in 6500+ records and data sets.

4.6.3 Proposed optimisation

During the initial Blue-Cloud project, the following approach was planned for retrieving from ELIXIR-ENA only records related to the marine environment. EBI should design one or more queries to constraint the whole set of studies (at that time 1040 records). A number of options have been investigated, with the following four conditions evaluated as being optimal.

1. inclusion in predefined projects such as Tara Oceans, Ocean Sampling Day, Malaspina, and others: projects of relevance include Tara Oceans (PRJEB402), Ocean Sampling Day (PRJEB5129) and Malaspina (PRJNA330770)
2. sampling attributed to aquatic environments: filters for marine- and aquatic-related text strings should be applied to “isolation_source”, “environment (biome)” and to values from tax_tree(410657)
3. sampling attributed marine region
4. organisms known to be aquatic

At the time, conditions 1-2 were directly available from the services. Therefore, the federation for ELIXIR-ENA has focused so far on these 2 conditions.

In the first part of the successor Blue-Cloud 2026 project, EBI has made further steps towards achieving a situation that more projects and related studies could be queried and harvested from ELIXIR-ENA.

For one, EBI has adopted Elastic to improve the performance of the system. Furthermore, EBI has implemented a method for tagging ELIXIR-ENA records, not only for the marine domain, but also for the wider aquatic domain, including oceans, seas, estuaries, rivers, and lakes. The following tags have been added:

WoRMS:
A table is used from WoRMS that contains taxa with their respective ranks and environmental attributes.

If a rank is a genus, subgenus, section, subsection, species, subspecies, variety, subvariety, forma, or subforma, then next to the taxon itself, all its descendants get the environmental attributes too.

Maps:

Four different maps are used for the 4 regions:

1. marine: openstreetmap-water-polygons/water_polygons.shp,
2. coastal: longhurst_v4_2010/Longhurst_world_v4_2010.shp,
3. freshwater: global200ecoregions/g200_fw.shp,
4. terrestrial: openstreetmap-land-polygons/land_polygons.shp

GeoTools :

Use is made of GeoTools[17], an open source Java library to handle the geospatial data. At the beginning of the indexing, all the shapefiles are loaded, based on:

https://docs.geotools.org/stable/userguide/library/data/shape.html

An important thing to consider whilst loading is setting the coordinate reference system (CRS). EBI set the CRS to WGS84 using `DefaultGeographicCRS.WGS84` and not `CRS.decode("EPSG:4326")`. There are other options, but these are the most common mappings. There is a major difference, which is that the first has (lon, lat) order, the second (lat,lon).

When the indexing starts, the geographic location of the records is tested against the 4 shapefiles, the library has a `contains` function. The EBI approach is based on this:

https://docs.geotools.org/latest/userguide/library/jts/geometry.html

[17] *https://docs.geotools.org*

Classifications

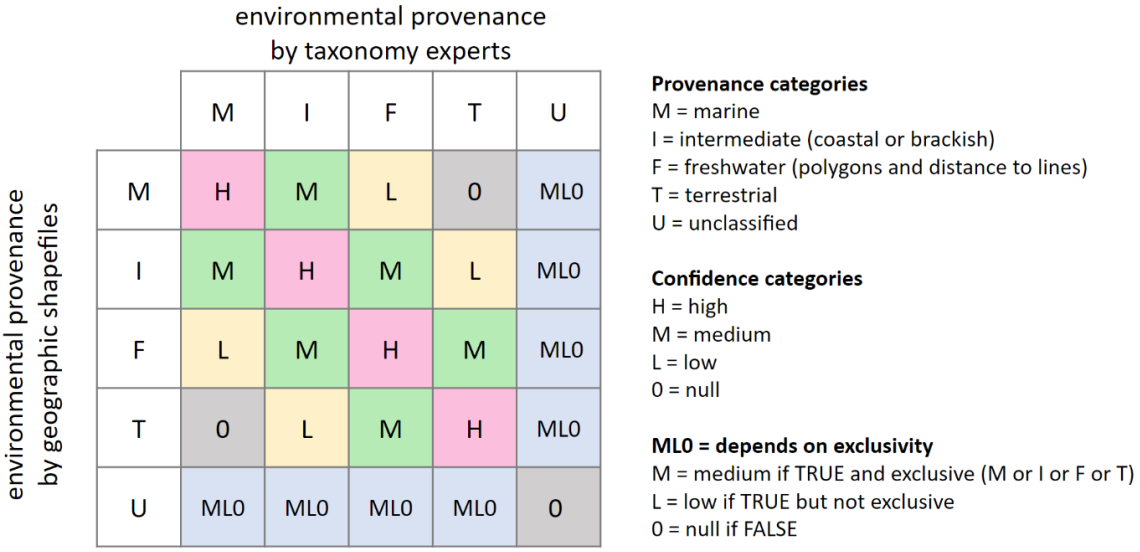

Figure 16 Matrix as followed for ELIXIR-ENA tagging to the aquatic projects and studies

The taxonomy and geographic information are combined based on a matrix which gives rise to the following classifications:

- marine:low_confidence
- marine:medium_confidence
- marine:high_confidence

- coastal_brackish:low_confidence
- coastal_brackish:medium_confidence
- coastal_brackish:high_confidence

- freshwater:low_confidence
- freshwater:medium_confidence
- freshwater:high_confidence

- terrestrial:low_confidence

- terrestrial:medium_confidence
- terrestrial:high_confidence

These tags are now available in ELIXIR-ENA for multiple data types:

- analysis
- sample
- read_run
- taxon

Note : the taxon does not have any geographic information and cannot get a higher confidence level than medium.

As a result, it is currently possible to query for data sets that are related to the aquatic environment and taxonomy. Previously, ELIXIR-ENA gave access in Blue-Cloud to circa 6500+ data records. The potential has now increased to > 1 million data sets as can be seen, when using the new tags for querying at **level 2** and giving counts of all records, for instance for the 'marine environment' tags:

```
curl --location 'https://www.ebi.ac.uk/ena/portal/api/count' \

      --header 'Content-Type: application/x-www-form-urlencoded' \

      --data-urlencode 'fields=all' \

      --data-urlencode 'result=read_study' \

      --data-urlencode 'format=json' \

      --data-urlencode 'query=(tag="marine")'
```

Overall, the ELIXIR-ENA currently provides the following counts :

- tag="freshwater": 1402669
- tag="freshwater:high_confidence": 12280
- tag="freshwater:medium_confidence": 42140
- tag="freshwater:low_confidence": 1346328
- tag="marine": 1303719
- tag="marine:high_confidence": 52755
- tag="marine:medium_confidence": 462981
- tag="marine:low_confidence": 745558

- tag="coastal_brackish": 927395
- tag="coastal_brackish:high_confidence": 12611
- tag="coastal_brackish:medium_confidence": 37752
- tag="coastal_brackish:low_confidence": 876745

It was asked whether use could be made of Elastic pagination for harvesting such large numbers of records for level 2, but EBI advices to query in 1 request to get all records at once. This is in their opinion the most efficient approach for the ENA API, although it concerns > 1 million records.

Adoption of this wider choice requires analysing how ELIXIR-ENA again can be approached in two levels, starting with level 1 focusing on study_accession records and followed by level 2 focusing on related samples and their data results.

The advanced API facilitates the level 2 queries for the different combinations in the tagging matrix. For keeping the Blue-Cloud DD&AS workflow intact, it is needed that EBI also includes a function in the API for retrieving a unique list of study_accession codes for each combination in the tagging matrix. That would allow to query and retrieve such a list for records with tags for e.g. ''marine:medium_confidence'' OR ''marine:high_confidence''. Such lists could then be used by CNR-IIA for retrieving the study_accesion metadata, to be included in the DAB service for level 1.

Another aspect is semantic interoperability. Most probably, ELIXIR-ENA makes use of a number of vocabularies for harmonising metadata and data descriptions. It would be useful if the use of vocabularies would be declared in the JSON and TSV output that the EBI ENA API by means of triples giving literal descriptions, term codes, and associated vocabulary urls. This enriched output then could be harvested and included in the Blue-Cloud DAB catalogue and OGC CSW service as well as used at level 2.

4.6.4 Planned actions

ACTION: To develop and expand the ENA API with functionality for listing unique study_accession codes for each combination in the new tagging matrix in support of the level 1 brokerage. To be undertaken by EBI.

ACTION: To indicate where and what vocabularies might being used in the ENA (meta)data model and to expand the JSON and TSV output from the ENA API with triples for literal description, coding of terms, and associated vocabulary URLs, where applicable, for optimising semantic interoperability. To be undertaken by EBI.

4.7. EcoTaxa data service

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, highthroughput 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.

4.7.1 Description

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.

Currently, EcoTaxa contains circa 340 million images of which circa 145 millions have been annotated in about 2000 projects, uploaded from ~650 organisations and classified by ~2200 users. Of these, circa 50% of 145 millions concern living organisms. The growth rate is circa 1 million images per month. Not all of these datasets are accessible for Blue-Cloud, because this depends on the data policy of data providers while EcoTaxa does not enforce datasets to be public.

4.7.2 Current federation

As part of the Blue-Cloud pilot project, It was decided to provide *summarised data* from EcoTaxa (concentrations per sample = lat/lon/time point) to EurOBIS, so that the Blue-Cloud broker can fetch the metadata about those datasets from EurOBIS. Blue-Cloud users should be able to discover the EcoTaxa datasets through EurOBIS and then, if more details are needed, turn to EcoTaxa to get them at granular (i.e. sample or object) level. Not all datasets in EcoTaxa will be uploaded to EurOBIS; but a selection of projects has been made which have resulted so far in **10 EurOBIS – EcoTaxa collections** which have been populated in EurOBIS. This amounts to several thousand data points containing approximately 5 million images with annotations. For populating its collections (i.e projects) in EurOBIS, EcoTaxa is using the IPT service. .

Level 1:

As a result, the DAB broker publishes EurOBIS collection records for EcoTaxa collections through the following endpoint:

https://blue-cloud.geodab.eu/gs-service/services/essi/view/ecotaxa/csw

with following GetCapabilities

https://blue-cloud.geodab.eu/gs-service/services/essi/view/ecotaxa/csw?service=CSW&request=GetCapabilities&version=2.0.2

The DAB Report on completeness can be found at:

https://dabreporting.s3.amazonaws.com/BlueCloud/BlueCloudReport_full.html

EcoTaxa available service at: **https://blue-cloud.geodab.eu/gs-service/services/essi/view/ecotaxa/csw**
EcoTaxa available test portal at: https://blue-cloud.geodab.eu/gs-service/search?view=ecotaxa
Total number of records: **10** Number of records analyzed: **10** Percentage of records analyzed: **100%**

Metadata element	Path	Completeness
IDENTIFIER	//gmd:fileIdentifier/gco:CharacterString	100%
TITLE	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:citation/gmd:CI_Citation/gmd:title/*[1]	100%
KEYWORD	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords[not(gmd:type) or not(contains('platform instrument',gmd:type/gmd:MD_KeywordTypeCode/@codeListValue))]/gmd:keyword/*[1]	100%
BOUNDING_BOX	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:geographicElement /gmd:EX_GeographicBoundingBox/gmd:westBoundLongitude/gco:Decimal /gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:geographicElement /gmd:EX_GeographicBoundingBox/gmd:eastBoundLongitude/gco:Decimal /gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:geographicElement /gmd:EX_GeographicBoundingBox/gmd:southBoundLatitude/gco:Decimal /gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:geographicElement /gmd:EX_GeographicBoundingBox/gmd:northBoundLatitude/gco:Decimal	100%
TEMPORAL_EXTENT	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:temporalElement/gmd:EX_TemporalExtent /gmd:extent/gml32:TimePeriod/gml32:beginPosition /gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:temporalElement/gmd:EX_TemporalExtent /gmd:extent/gml32:TimePeriod/gml32:endPosition	100%
PARAMETER	/gmi2019:MI_Metadata/gmd:contentInfo/gmi2019:MI_CoverageDescription/gmd:attributeDescription/gco:RecordType	100%
INSTRUMENT	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords[gmd:type /gmd:MD_KeywordTypeCode/@codeListValue='instrument']/gmd:keyword/*[1]	0%
PLATFORM	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords[gmd:type /gmd:MD_KeywordTypeCode/@codeListValue='platform']/gmd:keyword/*[1]	0%
ORGANIZATION	//gmd:CI_ResponsibleParty/gmd:organisationName/*[1]	100%
DATESTAMP	/gmi2019:MI_Metadata/gmd:dateStamp/gco:Date	100%
REVISION_DATE	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:citation/gmd:CI_Citation/gmd:date/gmd:CI_Date[gmd:dateType /gmd:CI_DateTypeCode/@codeListValue='revision']/gmd:date/gco:Date	100%
RESOURCE_IDENTIFIER	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:citation/gmd:CI_Citation/gmd:identifier/gmd:MD_Identifier/gmd:code /gco:CharacterString	100%

Figure 17 DAB completeness report of core metadata elements in the EcoTaxa service as determined from the EurOBIS web service

The output of the Blue-Cloud DAB broker is directly used to drive the level 1 discovery service of the Blue-Cloud Data Discovery & Access service.

Funded by the European Union

Blue-Cloud 2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227

Level 2:

As second level search, use is made of the **EcoTaxa Swagger API** to search for granular records inside a specific collection.

https://ecotaxa.obs-vlfr.fr/api/docs

For EcoTaxa, the natural granular records are objects and the best aggregation level for those granular records are samples; so ideally, it should facilitate Blue-Cloud users to search using the collection identifier, and a set of additional parameters, e.g.:

- spatial extent
- temporal extent
- depth extent
- free text in the name of the sample

The result set will contain matching samples, having as metadata elements the ones specified in the query, plus any other additional elements describing the granular record (e.g. title, id, platform, spatial extent, temporal extent, organization, ...). Having the sample id, the API then allows to easily query the object level information to export all the objects and images of these samples.

API results are harvested by MARIS at regular interval and queried from a local Elastic index. Reason for harvesting/caching at MARIS is that the API does not include an easy way to search/query the way that the DD&AS needs.

The data files are assembled by Maris from Elastic and consists of metadata. The 'download' product is the metadata itself and nothing more.

For level 2 the following search criteria are provided from the Elastic database:

- Free search
- Lat-Lon box
- Date period
- Classification
- Project

Criteria are provided as search facets, next to the free search, date search and geographic search. Currently, level 2 results in 6 Million+ records

4.7.3 Proposed optimisation

The number of EcoTaxa collections as populated in EurOBIS is currently 10 records. EcoTaxa is underway with expanding this in a number of ways :

- adding one dataset taken by a glider;
- enriching all existing datasets with concentration information in addition to occurrences. This required some work with EurOBIS to define how to format both occurrences AND concentration in the same DarwinCore Archive file. This is now done and the additional metadata should be collected by EurOBIS soon ;
- arranging that teams other than EcoTaxa operators can provide data collections to EurOBIS. This is currently held up by (1) the lack of a GUI in EcoTaxa to do this as biologists are not good at using APIs, (2) the fact that biologists and researchers in general are not very keen on spending time to share data, while preparing entries for EurOBIS requires efforts and following the format rules.

This results in a number of actions which are listed in the next paragraph.

Another aspect is improving the semantic interoperability of the EcoTaxa metadata. Most of the metadata fields in EcoTaxa have free text names provided by the users and not following any convention. A number of actions have been formulated to improve this situation by adopting controlled vocabularies and including declarations for these in the EcoTaxa metadata output.

A third aspect is enhancing options for querying data from EcoTaxa by subsetting. EcoTaxa's data is organised in hierarchical levels

- Project = cruise, time series
- Sample = station, time point
- Subsample = fraction of the original sample used for data acquisition
- Object = an image with its taxonomic identification

While collections = dynamic sets of projects (note : one project can be in several collections). The current flow in Blue-Cloud is:

- Collections are populated into EurOBIS
- DD&AS queries EurOBIS and gets the backlink to projects
- DD&AS queries all objects in the projects

At present, all object-level data is cached by the DD&AS. This implicates that DD&AS already can do subsetting as all required variables are available at object level. This works already in the current deployment of the DD&AS. It provides geolocalised occurence of the taxa, which are interesting for biogeography studies, but not concentrations, which are relevant for broader ecological studies. Therefore, a number of actions are formulated to explore options.

Funded by the European Union
Blue-Cloud 2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227

:

4.7.4 Planned actions

ACTIONS : to increase the flow of data to EurOBIS and Blue-Cloud DD&AS:

- **A.1. implement a GUI for the collections in EcoTaxa and their export as DarwinCore archive so that any EcoTaxa data owner can use it easily**
- **A.2 change the access policy to take the license into account: datasets with CC licenses will be queryable/downloadable by any logged in user. This will facilitate to harvest many more datasets from EcoTaxa.**

ACTIONS : to improve semantic interoperability:

- **C.1 EcoTaxa will allow users to pick which variables in their data match a set of variables from the Nerc Vocabulary Services (NVS) that are relevant for imaging datasets (defined through the EU JERICO-S3 project)**
- **C.2 EcoTaxa will implement import/export with a format that allows to define this match without manual intervention (likely the BioODV format)**
- **C.3 EcoTaxa will implement a general way of mapping any user-provided field to any NVS vocab term**

ACTIONS : to enhance the querying of data from EcoTaxa:

- **B.1: keep queries at object level, for occurrence data only: this requires EcoTaxa to implement an API endpoint (and probably an underlying database technology with possible support of MARIS) to allow slicing through the current 350M objects with acceptable performance**
- **B.2: add queries at (sub)sample level and add concentrations, for which there are options which have to further analysed by EcoTaxa and MARIS:**
  Step 1: EcoTaxa computes concentration of each taxon per (sub)sample; the DD&AS uses the "summary export" API endpoint to get those summarised tables for a whole project with the appropriate slicing variables; the DD&AS does the slicing. The machinery is there on EcoTaxa's backend, but the API end point and the UI to specify the formula to compute concentration do not exist.
  Alternative step 1: instead of querying the geolocalised concentrations from EcoTaxa, the DD&AS gets them from EurOBIS. All fields already exists at EurOBIS; no imaging dataset uses them yet but ~ all those provided during Blue-Cloud pilot soon will. Others might follow.
  Step 2: EcoTaxa adds endpoints to (1) search for samples according to the slicing variables, (2) get the summarised data for only those samples. Slicing is then done fully by EcoTaxa, not the DD&AS.

In terms of priority, EcoTaxa proposes : A => C.1 => B.1 => C.2 and possibly => B.2 and C.3
C.3

4.8. ICOS – Marine data service

ICOS ERIC 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.

4.8.1 Description

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.

4.8.2 Current federation

The ICOS Data portal (https://data.icos-cp.eu/portal/) uses a SPARQL endpoint based API underneath, that can be leveraged to harvest the desired collection metadata. The filters that should be set for marine observation data are the following:

- Project: ICOS
- Theme: Ocean data
- Data level: 1, 2

A correspondent SPARQL query can be executed through the ICOS SPARQL endpoint to retrieve the identifiers of the desired subset (with pagination):

```
# listFilteredDataObjects

prefix cpmeta: <http://meta.icos-cp.eu/ontologies/cpmeta/>

prefix prov: <http://www.w3.org/ns/prov#>

select ?dobj ?spec ?fileName ?size ?submTime ?timeStart ?timeEnd ?samplingHeight

where {

         VALUES       ?spec       {<http://meta.icos-cp.eu/resources/cpmeta/icosOtcL2Product>       <http://meta.icos-
cp.eu/resources/cpmeta/icosOtcL1Product> <http://meta.icos-cp.eu/resources/cpmeta/icosOtcL1Product_v2>}

          ?dobj cpmeta:hasObjectSpec ?spec .

          ?dobj cpmeta:hasSizeInBytes ?size .

          ?dobj cpmeta:hasName ?fileName .

          ?dobj cpmeta:wasSubmittedBy/prov:endedAtTime ?submTime .

          ?dobj cpmeta:hasStartTime | (cpmeta:wasAcquiredBy / prov:startedAtTime) ?timeStart .

          ?dobj cpmeta:hasEndTime | (cpmeta:wasAcquiredBy / prov:endedAtTime) ?timeEnd .

          FILTER NOT EXISTS {[] cpmeta:isNextVersionOf ?dobj}

}

order by desc(?submTime)

offset 0 limit 61
```

Records will be retrieved by the query. Then, each item can be queried with requests such as:

https://meta.icos-cp.eu/objects/OYFgMcIfy0zoH4M4EV-T6bQF?format=json

The returned JSON document describes each item with increased details. In the following table the mapping of some common metadata fields from the JSON response towards the Blue-Cloud metadata core elements is reported.

Blue-Cloud core metadata element	ICOS Data Portal metadata element
Identifier	$['PID']
Title	$['references'].['citationString']
Keyword	$['specification'].['project'].['keywords'] $['specificInfo].['acquisition'].['station].['org'].['name'] $['specificInfo].['columns'].['label'] project_name
Bounding box	$['coverageGeoJson'].['coordinates']

Blue-Cloud core metadata element	ICOS Data Portal metadata element
Temporal extent	$['specificInfo'].['acquisition'].['interval'].['start'] & $['specificInfo].['acquisition'].['interval'].['stop']
Parameter	$['specificInfo].['columns'].['label']
Instrument	Seems to be not available for this BDI
Platform	$['specificInfo].['acquisition'].['station].['org'].['name']
Organization	$['specificInfo].['productionInfo].['creator'].['creator'].['name']
Date stamp	$['specificInfo].['acquisition'].['interval'].['stop']
Revision date	$['specificInfo].['acquisition'].['interval'].['stop']

Table 8: mapping of ICOS Marine metadata elements to Blue-Cloud core elements

Level 1:

As a result, the DAB broker publishes ICOS Marine collection records through the following endpoint:

https://blue-cloud.geodab.eu/gs-service/services/essi/view/icos-data-portal/csw

with following GetCapabilities

https://blue-cloud.geodab.eu/gs-service/services/essi/view/icos-data-portal/csw?service=CSW&request=GetCapabilities&version=2.0.2

The DAB Report on completeness can be found at:

https://dabreporting.s3.amazonaws.com/BlueCloud/BlueCloudReport_full.html

ICOS Data Portal available service at: **https://blue-cloud.geodab.eu/gs-service/services/essi/view/icos-data-portal/csw**
ICOS Data Portal available test portal at: https://blue-cloud.geodab.eu/gs-service/search?view=icos-data-portal
Total number of records: **289** Number of records analyzed: **289** Percentage of records analyzed: **100%**

Metadata element	Path	Completeness
IDENTIFIER	//gmd:fileIdentifier/gco:CharacterString	100%
TITLE	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:citation/gmd:CI_Citation/gmd:title/*[1]	100%
KEYWORD	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords[not(gmd:type) or not(contains('platform instrument',gmd:type/gmd:MD_KeywordTypeCode/@codeListValue))]/gmd:keyword/*[1]	100%
BOUNDING_BOX	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:geographicElement /gmd:EX_GeographicBoundingBox/gmd:westBoundLongitude/gco:Decimal /gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:geographicElement /gmd:EX_GeographicBoundingBox/gmd:eastBoundLongitude/gco:Decimal /gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:geographicElement /gmd:EX_GeographicBoundingBox/gmd:southBoundLatitude/gco:Decimal /gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:geographicElement /gmd:EX_GeographicBoundingBox/gmd:northBoundLatitude/gco:Decimal	84%
TEMPORAL_EXTENT	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:temporalElement/gmd:EX_TemporalExtent /gmd:extent/gml32:TimePeriod/gml32:beginPosition /gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:temporalElement/gmd:EX_TemporalExtent /gmd:extent/gml32:TimePeriod/gml32:endPosition	100%
PARAMETER	/gmi2019:MI_Metadata/gmd:contentInfo/gmi2019:MI_CoverageDescription/gmd:attributeDescription/gco:RecordType	100%
INSTRUMENT	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords[gmd:type /gmd:MD_KeywordTypeCode/@codeListValue='instrument']/gmd:keyword/*[1]	12%
PLATFORM	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords[gmd:type /gmd:MD_KeywordTypeCode/@codeListValue='platform']/gmd:keyword/*[1]	100%
ORGANIZATION	//gmd:CI_ResponsibleParty/gmd:organisationName/*[1]	99%
DATESTAMP	/gmi2019:MI_Metadata/gmd:dateStamp/gco:Date	100%
REVISION_DATE	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:citation/gmd:CI_Citation/gmd:date/gmd:CI_Date[gmd:dateType /gmd:CI_DateTypeCode/@codeListValue='revision']/gmd:date/gco:Date	100%
RESOURCE_IDENTIFIER	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:citation/gmd:CI_Citation/gmd:identifier/gmd:MD_Identifier/gmd:code /gco:CharacterString	100%

Figure 18 DAB completeness report of core metadata elements in the ICOS Marine service as determined from their web service

The output of the Blue-Cloud DAB broker is directly used to drive the level 1 discovery service of the Blue-Cloud Data Discovery & Access service.

Level 2:

For level 2, use is made of the ERDDAP endpoint of ICOS-Marine:

https://erddap.icos-cp.eu

Results are harvested and included in a local Elastic Index, managed by MARIS, to support fast querying. The harvested set consists of all public datasets of the ICOS Ocean Thematic Centre, resulting in 250+ datasets.

An example of a data link:

https://erddap.icos-cp.eu/erddap/files/icos26na20170409SocatEnhanced/26NA20170409_SOCAT_enhanced.csv

For level 2 the following search criteria are provided from the Elastic database:

- Free search
- Lat-Lon box
- Date period
- Parameters

4.8.3 Proposed optimisation

It appears that the number of records as found through the SPARQL operation of the DAB broker at level 1 is circa 20 records higher than those found through the ERDDAP service. When asking around at ICOS, it became apparent that the ERDDAP service for ICOS-Marine is no longer maintained. It was only set-up for a pilot as part of the ENVRI-FAIR project. It is advised by ICOS to make only use of the SPARQL service as that is the sustained operational data discovery and access service of ICOS, which also includes the subset for ICOS-Marine. Therefore, the level 2 approach should be migrated from using ERDDAP to SPARQL. Moreover, retrieving data sets from the SPARQL service requires a user-password registration at ICOS.

Another aspect is improving the semantic interoperability of the ICOS Marine metadata. Further communication is needed with ICOS to see in how far use is made of controlled vocabularies, and if so, if these are already declared in the ICOS SPARQL output.

4.8.4 Planned actions

ACTION: To make use of the ICOS SPARQL service for level 2 and to register a Blue-Cloud login account with ICOS for allowing direct downloading. This will be undertaken by MARIS.

ACTION: To analyse with ICOS where and what vocabularies might being used in the ICOS (meta)data model, and if so, to discuss if triples for literal description, coding of terms, and associated vocabulary URLs, where applicable, could be included in the SPARQL service and output for optimising semantic interoperability. To be undertaken by MARIS and ICOS.

4.9. ICOS – SOCAT data service

4.9.1 Description

The Surface Ocean CO_2 Atlas (SOCAT) is a synthesis activity for quality-controlled, surface ocean fCO_2 (fugacity of carbon dioxide) observations by the international marine carbon research community. SOCAT data is publicly available, discoverable and citable. SOCAT enables quantification of the ocean carbon sink and ocean acidification and evaluation of ocean biogeochemical models. SOCAT represents a milestone in biogeochemical and climate research and in informing policy. SOCAT is a core Global Ocean Observing System data product for biogeochemistry endorsed by the Global Ocean Observing System (GOOS).

4.9.2 Current federation

The SOCAT portal (https://socat.info/) offers tools for interactively view the SOCAT data products and the download of synthesis and gridded files. This website is hosted by GeoMar and the Ocean Thematic Centre of ICOS.

The portal features an ERRDAP service which can be leveraged to harvest the metadata content of the different releases of SOCAT, for instance the 2020 release at:

https://ferret.pmel.noaa.gov/socat/erddap/tabledap/socat_v2020_fulldata.subset

So far, this 2020 version has been the focus of Blue-Cloud federation. It contans circa 6400+ records.

In the following table the mapping of some selected ERDDAP fields towards the Blue-Cloud metadata core elements is reported.

Blue-Cloud core metadata element	ICOS SOCAT metadata element
Identifier	expocode
Title	dataset_name
Keyword	platform_name
Bounding box	geospatial_lon_min & geospatial_lon_max & geospatial_lat_min & geospatial_lat_max &
Temporal extent	time_coverage_start & time_coverage_end
Parameter	Fixed values for all records: *salinity, sea surface temperature, sea-level air pressure, WOCE flag for aqueous CO2, fCO2*
Instrument	Seems to be not available for this BDI
Platform	platform_name

Blue-Cloud core metadata element	ICOS SOCAT metadata element
Organization	organization
Date stamp	time_coverage_end
Revision date	time_coverage_end

Table 9: mapping of SOCAT metadata elements to Blue-Cloud core elements

Level 1:

As a result, the DAB broker publishes EurOBIS collection records through the following endpoint:

https://blue-cloud.geodab.eu/gs-service/services/essi/view/icos-socat/csw

with following GetCapabilities

https://blue-cloud.geodab.eu/gs-service/services/essi/view/icos-socat/csw?service=CSW&request=GetCapabilities&version=2.0.2

The DAB Report on completeness can be found at:

https://dabreporting.s3.amazonaws.com/BlueCloud/BlueCloudReport_full.html

ICOS SOCAT available service at: **https://blue-cloud.geodab.eu/gs-service/services/essi/view/icos-socat/csw**
ICOS SOCAT available test portal at: https://blue-cloud.geodab.eu/gs-service/search?view=icos-socat
Total number of records: **7484** Number of records analyzed: **7484** Percentage of records analyzed: **100%**

Metadata element	Path	Completeness
IDENTIFIER	//gmd:fileIdentifier/gco:CharacterString	100%
TITLE	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:citation/gmd:CI_Citation/gmd:title/*[1]	43%
KEYWORD	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords[not(gmd:type) or not(contains('platform instrument',gmd:type/gmd:MD_KeywordTypeCode/@codeListValue))]/gmd:keyword/*[1]	100%
BOUNDING_BOX	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:geographicElement /gmd:EX_GeographicBoundingBox/gmd:westBoundLongitude/gco:Decimal /gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:geographicElement /gmd:EX_GeographicBoundingBox/gmd:eastBoundLongitude/gco:Decimal /gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:geographicElement /gmd:EX_GeographicBoundingBox/gmd:southBoundLatitude/gco:Decimal /gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:geographicElement /gmd:EX_GeographicBoundingBox/gmd:northBoundLatitude/gco:Decimal	91%
TEMPORAL_EXTENT	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:temporalElement/gmd:EX_TemporalExtent /gmd:extent/gml32:TimePeriod/gml32:beginPosition /gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:extent/gmd:EX_Extent/gmd:temporalElement/gmd:EX_TemporalExtent /gmd:extent/gml32:TimePeriod/gml32:endPosition	91%
PARAMETER	/gmi2019:MI_Metadata/gmd:contentInfo/gmi2019:MI_CoverageDescription/gmd:attributeDescription/gco:RecordType	100%
INSTRUMENT	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords[gmd:type /gmd:MD_KeywordTypeCode/@codeListValue='instrument']/gmd:keyword/*[1]	0%
PLATFORM	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords[gmd:type /gmd:MD_KeywordTypeCode/@codeListValue='platform']/gmd:keyword/*[1]	100%
ORGANIZATION	//gmd:CI_ResponsibleParty/gmd:organisationName/*[1]	100%
DATESTAMP	/gmi2019:MI_Metadata/gmd:dateStamp/gco:Date	100%
REVISION_DATE	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:citation/gmd:CI_Citation/gmd:date/gmd:CI_Date[gmd:dateType /gmd:CI_DateTypeCode/@codeListValue='revision']/gmd:date/gco:Date	91%
RESOURCE_IDENTIFIER	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:citation/gmd:CI_Citation/gmd:identifier/gmd:MD_Identifier/gmd:code /gco:CharacterString	100%

Figure 19 DAB completeness report of core metadata elements in the SOCAT service as determined from their web service

The output of the Blue-Cloud DAB broker is directly used to drive the level 1 discovery service of the Blue-Cloud Data Discovery & Access service.

Level 2:

For level 2, also use is made of the ERDDAP endpoint of SOCAT:

https://ferret.pmel.noaa.gov/socat/erddap/tabledap/socat_v2020_fulldata.subset

Results are harvested and included in a local Elastic Index, managed by MARIS, to support fast querying. The harvested set consists of all public datasets of SOCAT, resulting in 6400+ datasets.

An example of a data link:

https://data.pmel.noaa.gov/socat/erddap/files/socat_v2020_fulldata /3164/316420110205.nc

For level 2 the following search criteria are provided from the Elastic database:

- Free search
- Lat-Lon box
- Date period
- Platform Type
- Platform Name
- QFlag
- Investigators
- Organisation

The fields are made available as facets next to the free search, date and geospatial search criteria.

4.9.3 Proposed optimisation

There are multiple versions of SOCAT. Currently, the DD&AS has federated the 2020 version, but recently, SOCAT has released the 2023 version. This version can be downloaded as a TSV file, while it is also made available by NOAA through their ERDDAP service at :

https://data.pmel.noaa.gov/socat/erddap/tabledap/socat_v2023_fulldata.subset

When comparing the number of ExpoCodes included in the TSV and ERDDAP service, there are differences. After checking with SOCAT and NOAA, it appears that the TSV version is filtered for data of bad quality, while these are still included in the ERDDAP service. The ERDDAP includes quality flags, which

allows then to filter also the ERRDAP output that will be federated in the DD&AS. In particular, it concerns data that is Questionable which is indicated with records having the WOCE flag 3.

Currently, at level 1 for most records there is no TITLE available. This could be corrected by completing the TITLE field with the ExpoCode (Source Identifier), which will give 100% population. Expocodes represent uniq events of measurements made on ships, moorings, autonomous and drifting surface platforms for the global ocean and coastal seas.

Another aspect is improving the semantic interoperability of the SOCAT metadata. A check on ERDDAP reveals that there are no vocabulary declarations included. Further communication is needed with SOCAT to see in how far use is made of controlled vocabularies, and if so, if these could be declared as triples in the SOCAT output.

4.9.4 Planned actions

ACTION: To replace the 2020 version from SOCAT ERDDAP with the 2023 version, and filtering for the records for which included quality flags are not all 'bad'. This will be undertaken by CNR-IIA for level 1 and MARIS for level 2.

ACTION : To replace filling of the TITLE field at level 1 from 'Datasetname' to 'Expocode' to get a 100% filling. To be undertaken by CNR-IIA.

ACTION: To analyse with SOCAT where and what vocabularies might being used in the SOCAT (meta)data model, and if so, to discuss if triples for literal description, coding of terms, and associated vocabulary URLs, where applicable, could be included in the ERDDAP service and output for optimising semantic interoperability. To be undertaken by MARIS and SOCAT.

5. General optimisation options

The previous chapter gives optimisations for the Blue-Cloud Data Discovery & Access Service (DD&AS) which are aimed at the federated Blue Data Infrastructures. Next to these, there are also options for optimisations of the central components of DD&AS.

5.1. Semantic brokerage in support of Data Discovery

5.1.1 Introduction

In the current DD&AS approach, there are two levels for searching :

- Level 1 : at collection level with a few general operators (free search, geospatial search and temporal search) to identify which of the BDIs might have interesting data sets
- Level 2 : at granular level with a search profile per BDI to drill down in searching and to get access to the identified data sets by downloading, supported by a common shopping basket and delivery dashboard

In the previous chapter, it is described that already most of the BDIs are making use of controlled vocabularies and actions have been formulated to make this more explicit in the metadata output by means of including triples with literal description, code term, and associated vocabulary. This enriched output will then allow the DD&AS to include the semantic information also in the operations at level 1 and level 2. For some BDIs some further checking is needed, but overall the situation towards semantic info provision looks promising.

By having this semantic information in the DD&AS metadata records, then should facilitate to make use of this in the queries and in the presentations of the output.

Considering the present situation, it is expected that several BDIs make use of the Nerc Vocabulary Services, WoRMS, and EDMO (organisations), while also other vocabularies might be in use. The actions for semantic interoperability will allow to make a matrix, which will give an overview of fields per BDI at level 1 and level 2, which are supported by a controlled vocabulary and which vocabulary.

Having that matrix will facilitate looking into semantic brokering between different vocabularies in use for comparable metadata fields, such as parameters, platforms, instruments, organisations, sea regions, etc.

And having such a semantic brokerage would then allow to expand the search criteria already at level 1, using more of the fields of the common Blue-Cloud DAB metadata model. In total 13 metadata elements from ISO 19115 are considered as the common elements of the Blue-Cloud DAB profile. These common

Blue-Cloud metadata elements are given below and the ones, potentially supported by controlled vocabularies, are highlighted:

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

The current level 1 search only includes : Free Search, Temporal Search ; Geospatial Search. Additional highlighted terms might be added as facet search, making the lefel 1 search more precise and powerful.

Pre-conditions for this expansion of the level 1 search profile are :

- BDIs should make use of controlled vocabularies for candidate search criteria
- BDIs should have content in their databases to cover these candidate search criteria
- DD&AS should have a semantic brokerage service to make cross walks between different vocabularies used for specific candidate search criteria.

For current level 2, search profiles are different per BDI, but for candidate metadata fields, that fulfil the pre-conditions, again use could be made in their search profiles, providing harmonisation for commonly used terms.

Finally, the semantic information and brokerage could also be used for harmonisation of common fields in the detail information at published for the BDI record. This could be applied at level 1 and level 2 output.

5.1.2 How to make use and set up semantic brokering

The Nerc Vocabulary Service (NVS) is used by many organisations and infrastructures worldwide and it belongs to the SeaDataNet standards and services which are widely promoted by SeaDataNet in the European marine data management community to achieve standardisation in metadata, data and data

products. The NVS service is operated and managed by NOC-BODC, who have built up great expertise in this domain of semantic interoperability.

From NOC-BODC experience, instruments and platforms in the marine domain would benefit from being harmonised against the standard set up by SeaDataNet, and already adopted by many marine organisations worldwide.

Instruments

The following controlled vocabularies are used to form an instrument scheme or thesaurus: https://vocab.nerc.ac.uk/scheme/SDNDEV/current/

Where top level = https://vocab.nerc.ac.uk/collection/L21/

Followed by https://vocab.nerc.ac.uk/collection/L05/ for instrument types

And by https://vocab.nerc.ac.uk/collection/L22/ for instrument models

So by using L22 in their data models for their instrument model identification, BDIs can gain access and make use of the links established on the NVS to additional concepts like instrument types L05 and L21 and also link to the instruments manufacturers via https://vocab.nerc.ac.uk/collection/L35/

For example, for ECOTAXA, a number of entries match some of the instruments mentioned in the metadata records:

“Uvp5” can be matched to the following L22 concepts: https://vocab.nerc.ac.uk/collection/L22/current/TOOL1577/, https://vocab.nerc.ac.uk/collection/L22/current/TOOL1650/,

https://vocab.nerc.ac.uk/collection/L22/current/TOOL1578/

While “zooscan” can be matched to:

https://vocab.nerc.ac.uk/collection/L22/current/TOOL1581/

Their suitability could be evaluated as part of the project.

Platforms

For platforms, SeaDataNet and associated infrastructures already use controlled vocabularies for :

Funded by the European Union
Blue-Cloud 2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union’s Horizon Europe programme under Grant Agreement no. 101094227

- platform categories : https://vocab.nerc.ac.uk/collection/L06/
- platform models : https://vocab.nerc.ac.uk/collection/
- platform instances : sourced from ICES or https://vocab.nerc.ac.uk/collection/C17/current/

So, for example, SOCAT Platform types “Ship”, “Mooring”, “Drifting buoy”, “Autonomous Surface Vehicle”, “Boat” align to:

https://vocab.nerc.ac.uk/collection/L06/current/30/
https://vocab.nerc.ac.uk/collection/L06/current/48/
https://vocab.nerc.ac.uk/collection/L06/current/42/
https://vocab.nerc.ac.uk/collection/L06/current/3B/

The fifth term “Boat” could map to several concepts in L06 and would therefore need more explanations from the BDI before applying a mapping. I.e. what is meant by “boat”:
https://vocab.nerc.ac.uk/collection/L06/current/33/?
https://vocab.nerc.ac.uk/collection/L06/current/37/?
https://vocab.nerc.ac.uk/collection/L06/current/3A/?
https://vocab.nerc.ac.uk/collection/L06/current/38/?

This exercise can be repeated for each BDI and we could obtain alignment on those 2 fields Instruments and Platforms.

Keywords

Additionally, keywords could be matched to their category in :

https://vocab.nerc.ac.uk/collection/L19/current/

further helping the semantic alignment and powering semantic search on key metadata components, down to the variable categories.

Variables/parameters

As part of the EU ENVRI-FAIR project, NOC-BODC developed the EOV demonstrator that introduced semantic search on variables. For these concepts were used from the “Essential Ocean Variables” family as a common source of high level variable “categories”. A pre-existing NVS collection called A05 was used as a source of concepts that users could choose from (the demonstrator only focused on a subset of them).

Once selected, the concept needs to enable discovery of variables held in datafiles. For this, either needed a **direct connection** between the source and target concepts or a **crosswalk** is used to go from the source to the target concepts. The former uses traditional direct broad-narrow mappings which results in large amounts of mappings, and complex and costly maintenance. The latter was chosen and use was made of the **I-ADOPT ontology**[18] as the connector between the broad discovery term and the more detailed parameter codes held in the files (e.g. BODC P01 codes or CF Standard Names). This is referred to as "smart mappings".

For example, for ENVRI-FAIR NOC-BODC implemented the following connections:

- A05 concepts mapped to I-ADOPT key atomic components defined in reference controlled vocabularies
- P01 concepts mapped to I-ADOPT key atomic components using the same reference vocabularies
- The connection between A05 and P01 concepts then gives direct access to any other concept linked to the P01: e.g. P02 (used by SeaDataNet at the CDIs parameter discovery level, R03 used by Argo for parameter identification, P09 used by other networks, CF Standard Names used in CF netCDF compliant files)
- These relationships are then used by the semantic broker via SPARQL query to fetch results that match the alignment with the selected search A05 concept

As candidate vocabularies for level 1 search related to variables, we would recommend using simple vocabularies that only have a small number of broad concepts: it could be EV concepts, terms related to environmental monitoring like e.g. P36 concepts, but also high level disciplines like e.g. P08. The suitability of other vocabularies could also be evaluated.

Organisations

For organisations it is recommended to make use of the SeaDataNet EDMO (Directory of Organisations) which already is used by many infrastructures worldwide and for which there is an active organisation for maintaining existing entries and processing new EDMO requests. The later applies in practice also for the NVS vocabularies, which have an established governance organisation and are steadily expanding in number of terms and types of vocabularies, following the needs from the marine communities.

[18] https://i-adopt.github.io/

EDMO can be found at :

https://edmo.seadatanet.org/

And like NVS it features a GUI, a SOAP web service, and a SPARQL endpoint. EDMO is centrally operated by MARIS as one of the SeaDataNet resources, with SeaDataNet partners as national nodes for maintenance of contents, and also some international accounts in USA (Scripps) and Australia (IMOS), following the EU Ocean Data Interoperabilty Platform (ODIP[19]) projects. EDMO not only gives a PID, full addresses, and abstracts, but also links to roles performed by the organisations in other SeaDataNet directories, such as EDMED (Data sets), CDI (Common Data Index), EDMERP (Research projects), EDIOS (Observing programmes, networks and stations), and CSR (Cruise Summary Reports).

For instance, the EDMO landing page for Ifremer can be found at :

https://edmo.seadatanet.org/report/486

5.1.3 How to make use and set up semantic brokering

ACTION: To make a matrix of vocabularies in use by each BDI for common metadata fields such as parameters, instruments, platforms, keywords, organisations, and possible others, following the semantic actions as formulated in Chapter 4 by BDIs. To review where mappings or crosswalks between vocabularies for same fields could be applied and situations where BDIs might adopt vocabularies, recommending SeaDataNet vocabularies (NVS + EDMO) as reference vocabularies. To be undertaken by MARIS, NOC-BODC, CNR-IIA and BDIs.

ACTION: To provide support to BDIs for uptake of vocabularies, also considering required mapping per BDI of current terms to SDN vocab terms, and for incorporating and producing the 'triples' in the BDI metadata. To be undertaken by MARIS, NOC-BODC, CNR-IIA, and BDIs.

[19] https://www.odip.org

ACTION: To develop a semantic brokerage mechanism, using mappings and I-ADOPT framework, foreseeing that the planned semantic provision by BDIs will be established. To integrate this mechanism in the DD&AS at level 1 and 2. To be undertaken by NOC-BODC, MARIS, and CNR-IIA.

5.2. Extra functionality of data subsetting

Currently, the DD&AS supports discovery and download of predefined data objects, which are documented with metadata records. Using the metadata, interesting data sets can be discovered and identified, and related data sets can be retrieved following the download information. This is a very useful functionality.

In addition, there are applications and use cases which would like to have a sub-setting or slicing functionality, which works at the level of the data sets, and would facilitate to extract from the data sets for specific parameters and values, for instance all temperature observations with temperature value > 10 degrees in October 2020 between 50 and 100 meter depth. Such a functionality could also be combined with the already established discovery and download functionality of the DD&AS.

Such an additional functionality will be very useful and efficient for a number of the Blue-Cloud Virtual Labs and Work Benches as it will allow to build up and maintain data lakes for specific parameters by extracting regularly from BDI repositories.

During the TSC meeting in Amsterdam, there was a general interest in further exploring and deploying such sub-setting functionality. Also, during the bilateral meetings with each of the BDIs, there was a clear support for analysing and implementing such functionality by means of APIs at the BDIs. Some BDIs are managing data, which in fact is metadata, such as EcoTaxa and EBI-ENA, and subsetting is there already functioning as part of the use of Elastic for indexing and querying. Other BDIs are managing data sets which include a lot numerical data and which are provided in formats such as NetCDF, ODV ASCII, and other ASCII formats.

There are some tools available, that support sub-setting, such as the **THREDDS Data Server** (*TDS*)[20], based

[20] https://www.unidata.ucar.edu/software/tds/

on **OPeNDAP**[21], **ERDDAP**[22], and others. A promising development in this field is being developed by MARIS, the **BEACON** tool.

Access to a large number of multidisciplinary data resources is key. However, achieving performance is a major challenge as original data is organized in millions of observation files which makes it hard to achieve fast responses. Next to this, data from different domains are stored in a large variety of data infrastructures, each with their own data-access mechanisms, which causes researchers to spend much time on trying to access relevant data. In a perfect world, users should be able to retrieve data in a uniform way from different data infrastructures following their selection criteria, including for example spatial or temporal boundaries, parameter types, depth ranges and other filters.

Looking at this perspective, MARIS is developing the BEACON software system with a unique indexing system that can, on the fly, extract specific data based on the user's request from millions of observational datafiles containing multiple parameters in diverse units. The BEACON system and its data can be accessed via a REST API that is exposed by BEACON itself meaning clients can query data via a simple JSON request. The system is built in a way that it returns one single harmonized file as output, regardless of whether the input contains many different datatypes or dimensions. It also allows for converting the units of the original data if parameters are measured in different types of units. It is important to mention that the system can be applied to different data infrastructures and is not tailor made for one specific type of database. Currently. MARIS is testing the BEACON API for the SeaDataNet CDI database, the ERA5 dataset from the Climate Data Store, and the full Argo data set, to showcase its performance and user friendliness. Preliminary results are very good and demonstrations given to BDIs and at the recent EuroGOOS 2023 Conference have created interest from several infrastructures, including several Blue-Cloud BDIs to see what BEACON could mean for their services.

ACTION : The topic of expanding the DD&AS functionality with sub-setting services and developing specifications for Blue-Cloud data lakes is well underway and its analysis results will be documented in Deliverable D2.4 - BDI sub-setting APIs and Data Lakes – Concept and Specifications Report, which is planned for end February 2024.

[21] https://www.opendap.org/
[22] https://github.com/ERDDAP/erddap

5.3. Replace Marine-ID for federated EOSC AAI

Currently, the Blue-Cloud DD&AS makes use of the SeaDataNet Marine-ID service for user registration and AAI services. All data in the DD&AS concern open data with CC-BY-4.0 license. But users need to be registered to be able to submit data shopping requests and to follow their transactions through their personal dashboard.

As part of EOSC, there is a lot of focus on supporting federated AAI, which can be achieved by making use for logon of EGI-Checkin.

https://www.egi.eu/service/check-in/

This service allows users, registered in many other AAI services, to logon with their original logon details to many service providers. EGI Check-in gives a centralised solution that seamlessly connects various Identity Providers with EGI service providers. It facilitates a user to quickly and easily select its preferred Identity Provider and gain access to a wide range of services without any hassle.

The current Marine-ID service is not fit for such a federation and is already quite old, while there are now more modern AAI services available which are better fit for such federations. A good alternative is KeyCloak, which will be deployed by MARIS as Blue-Cloud DD&AS AAI service, and which can be easily federated in EGI-Checkin. A comparable migration has been established earlier by CNR-ISTI for giving logon access to the Blue-Cloud Virtual Research Environment.

ACTION: MARIS will deploy an instance of KeyCloak to replace the current Marine-ID AAI service as use in the Blue-Cloud DD&AS, which will be connected to the EGI-Checkin. To be undertaken by MARIS.

5.4. Extra monitoring of federation

The Blue-Cloud DD&AS is striving for TRL7 or more, which implicates an operational and robust service. The DD&AS is largely based upon machine-to-machine interactions with web services and APIs from multiple BDIs. This makes the DD&AS operationally vulnerable for disturbances at BDIs for those web services. Partly this is compensated by pre-procesing of the level 1 records, currently at a weekly harvest schedule, and which is automatically repeated in case of temporary glitches at BDI level. For level 2, a

comparable caching situation is available, at least for half of the number of BDIs. While for the data retrieval, there is dependence on the BDIs for almost all of them.

Having a good insight in the operational availability of all components of the DD&AS will allow to identify weak nodes and to undertake action for improvement.

ACTION: To explore and analyse how an operational availability monitoring of DD&AS service components could be established, e.g. by making use of the EOSC core service for monitoring. To be undertaken by MARIS with contributions from CNR-IIA and BDIs.

5.5. Developing API for DD&AS

ACTION: To develop a Swagger API version of the Blue-Cloud DD&AS, once the optimisations have been successfully completed. To be undertaken by MARIS.

5.6 Expanding the DD&AS with additional BDIs

Additional BDIs will be connected to the DD&AS federated service. These concern: EMODnet Physics, ELIXIR-MGnify, SIOS, and EMSO. Analyses for connecting these BDIs are ongoing.

ACTION: To expand the Blue-Cloud DD&AS with additional BDIs. A first analysis will be documented in D2.2 - New Blue Data Infrastructures – Service Analysis Report, which is planned for end October 2023. To be undertaken by MARIS with new BDIs (EMODnet Physics, ELIXIR-MGnify, SIOS, and EMSO.

6. Conclusions and follow-up

The D2.1 report gives an overview of all the optimisations that are planned for improving the functioning and services of the Blue-Cloud Data Discovery & Access Service (DD&AS). The developments will concern activities at the level of the federated Blue Data Infrastructures (BDIs) and at the central components of the DD&AS. The report lists all foreseen actions for achieving the overall optimisation.

These actions are partly ongoing and partly planned in the coming period. Their progress and results will be documented in the following Deliverables, listed on time :

- D2.2 : New Blue Data Infrastructures – Service Analysis Report – end Oct 2023
- D2.4 : BDI sub-setting APIs and Data Lakes – Concept and Specifications Report – end Feb 2024
- D2.6 : Tuning between Blue-Cloud Data Lakes and DTO development, 1st report – end Feb 2024
- D2.3 : Optimised and expanded Blue Cloud Data Discovery and Access Service – Documentation Report – end Dec 2024
- D2.5 : Established BDI sub-setting APIs and Data Lakes – Documentation Report – end Feb 2025
- D2.7 : Tuning between Blue-Cloud Data Lakes and DTO development, 2nd report – end Feb 2025

Funded by the European Union
Blue-Cloud 2026 – A federated European FAIR and Open Research Ecosystem for oceans, seas, coastal and inland waters has received funding from the European Union's Horizon Europe programme under Grant Agreement no. 101094227