Blue-Cloud

Piloting innovative services for Marine Research & the Blue Economy

D2.4 Blue-Cloud Data Discovery and Access service

Work Package	WP2, developing the Blue Cloud discovery and access service and overall Blue Cloud architecture
Lead Partner	MARIS
Lead Author (Org)	MARIS, CNR-IIA, CINECA, IFREMER
Contributing Author(s)	Dick M.A. Schaap (MARIS), Enrico Boldrini (CNR-IIA), Gilbert Maudire (IFREMER), Mattia D'Antonio (CINECA)
Reviewers	TCOM members*
Due Date	28.02.2021, M17 [Extension agreed with EC to 31 May 2021]
Submission Date	31.05.2021
Version	1.0

Dissemination Level

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

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

DISCLAIMER

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

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

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

COPYRIGHT NOTICE

VERSIONING AND CONTRIBUTION HISTORY

Version	*Date*	*Authors*	*Notes*
0.1	31.05.2021	MARIS	First version
1.0	18.11.2020	TRUST-IT	Deliverable submission

**The current version 1.0 of this deliverable still have to pass the internal review process. The authors and the project coordinators decided to submit the deliverable prior to the completion of the internal review process in order to make the document available for the external reviewers appointed by the European Commission for the first interim project review, scheduled on 21st and 22nd of June*

Contents

Executive summary 5

1 Introduction 6

2 Overall concept of Blue Cloud data discovery and access service 7

2.1 Overall concept 7

2.2 Involved blue data infrastructures 9

2.3 Architecture and modules 12

3 Level 1 – Common Blue-Cloud metadata catalogue at collection level 14

3.1 CSW for ISO Blue-Cloud profiles per blue data infrastructure 15

3.2 Euro Argo – ARGO GDAC service 18

3.3 ELIXIR-ENA service 20

3.4 EMODnet Chemistry Data Products service 22

3.5 EurOBIS – EMODnet Biology service 24

3.6 ICOS Marine Data Portal service 26

3.7 ICOS SOCAT service 28

3.8 SeaDataNet CDI service 30

3.9 SeaDataNet Data Products service 32

3.10 EcoTaxa service 34

3.11 Blue-Cloud first level (collection) metadata report – internal service 35

4. Level 2 – Individual search profiles at granule level 37

4.1 Euro Argo – ARGO GDAC service 38

4.2 ELIXIR-ENA service 39

4.3 EMODnet Chemistry Data Products service 39

4.4 EurOBIS – EMODnet Biology service 40

4.5 ICOS Marine Data Portal service 41

4.6 ICOS SOCAT service 41

4.7 SeaDataNet CDI service 42

4.8 SeaDataNet Data Products service 42

4.9 EcoTaxa service 43

5. Delivery service 44

6. Deployment of the Blue-Cloud Data Discovery and Access service 46

Glossary

Term	Definition
APIs	Application Programming Interface
BDI	Blue Data Infrastructures
CDI	Common Data Index
CMEMS	Copernicus Marine Environmental Monitoring Service
CSW	
DAB	
DOI	Digital Object Identifier
EEA	European Environment Agency
EMBL-EBI	EMBL European Bioinformatics Institute
EMODnet	European Marine Observation and Data Network
ENA	European Nucleotide Archive
ERDDAP	Environmental Research Division Data Access Program
GB	Gigabyte
GDAC	Global Data Assembly Centres
GBIF	Global Biodiversity Information Facility
GUI	
LOV	Laboratoire d'Océanographie de Villefranche
HTTP-GET	
HTTP-POST	
ICES	International Council for the Exploration of the Sea
ICOS	European Research Infrastructure Integrated Carbon Observation System
ISO	International Organization for Standardization
JSON	JavaScript Object Notation
MarBEF	MARine Biodiversity and Ecosystem Functioning
NODCs	National Oceanographic Data Centres
OBIS	Ocean Biogeographic Information System
OGC	Open Geospatial Consortium
REST	Representational State Transfer
SOCAT	Surface Ocean CO_2 Atlas
SQL	Structured Query Language
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
TB	Terabyte
WMTS	Web Map Tile Service
WWWPIC	World Wide Web of Plankton Image Curation
VRE	Virtual Research Environment
XML	Extensible Markup Language

Executive summary

The **Blue Cloud data discovery and access service** is one of the components of the Blue-Cloud technical framework. It serves federated discovery and access to a range of blue data infrastructures for external users and also will interact with the Blue-Cloud Virtual Research Environment (the component federating computing platforms and analytical services). The pilot Blue-Cloud project aims at federating initially in total 10 blue data infrastructures. Each of these existing infrastructures have been described earlier in deliverable D2.1 - Blue Data Infrastructures – Services Description Report, in particular with a focus on their current data discovery and access mechanisms. And further analyses have taken place concerning the web services and APIs that each of the existing infrastructures are providing for machine-to-machine interactions. The results of those analyses have been described in deliverable D2.2 - Blue Data Infrastructures – Services Analysis Report.

The conceptual approach and architecture for the Blue Cloud data discovery and access service have been further developed and are described in deliverables D2.6 – Blue Cloud Architecture (1st Release) and D2.7 – Blue Cloud Architecture (2nd Release). Part of the concept is to rely fully on machine-to-machine interactions between the Blue-Cloud service and each of the blue data infrastructures. Another part of the concept is a two-step approach to discovery from data collections to data granules, where possible, and applying search criteria from coarse (level 1) to fine (level 2), while the second level can be completed by actual ordering and retrieval of associated data sets.

The underlying report describes the actual development and deployment of the Beta-version of the Blue-Cloud Data Discovery and Access service. Implementing the conceptual approach has largely depended on the existing and available web services and APIs at the blue data infrastructures. Therefore, this report describes what services and what solutions were adopted overall as well as per blue data infrastructure to deploy the requested concept and functionality.

The actual launch of the Beta-version of the Blue-Cloud Data Discovery and Access service is planned in the middle of June 2021.

1 Introduction

The **Blue Cloud Data Discovery and Access service** initially has been planned to serve federated discovery and access to the following blue data infrastructures:

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

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

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

Activities in Task 2.1 have resulted in deliverable D2.1 - Blue Data Infrastructures – Services Description Report, which describes each of the blue data infrastructures and in particular their current data discovery and access mechanisms. While activities in Task 2.2 have resulted in deliverables D2.6 – Blue Cloud Architecture (1st Release) and D2.7 – Blue Cloud Architecture (2nd Release), which describe with new insights the architecture and concept for the Blue Cloud data discovery and access service. The implementation of the Blue Cloud data discovery and access service will largely depend on the machine-to-machine interfaces of the blue data infrastructures. For that purpose, further analysis activities as part of Task 2.2 have been undertaken by MARIS and CNR-IIA, interacting with technical experts of the blue data infrastructures, concerning the functioning of their existing web services, and in how far these are already fit for purpose or require further developments. These findings are described in deliverable D2.2 - Blue Data Infrastructures – Services Analysis Report.

Very good progress has been made with the actual development and deployment of the Blue Cloud Data Discovery and Access service, so that a launch of the Beta-version is now planned for the middle of June 2021.

The underlying report will describe what services and what solutions were adopted overall as well as per blue data infrastructure to deploy the Blue Cloud Data Discovery and Access service, taking into account the overall concept and the variety in web services, APIs, and formats provided by the different blue data infrastructures.

2 Overall concept of Blue Cloud data discovery and access service

2.1 Overall concept

The **Blue-Cloud Data Discovery and Access service** is one of the two main components of the Blue-Cloud technical framework, next to the **Blue Cloud Virtual Research Environment (VRE)**. The Blue Cloud service will facilitate 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 that will be connected to the Blue Cloud service to serve federated discovery and access. The development of the Blue Cloud Data Discovery and Access service is undertaken as part of WP2.

The pilot Blue-Cloud Data Discovery and Access service project aims at federating initially in total 10 blue data infrastructures. Each of these existing infrastructures have been described in deliverable D2.1 - Blue Data Infrastructures – Services Description Report, in particular with a focus on their current content and data discovery and access mechanisms. While deliverable D2.2 - Blue Data Infrastructures – Services Analyses Report – describes the web services and APIs that each of the blue data infrastructures is providing and maintaining.

The overall concept is that the Blue-Cloud Data Discovery and Access service will make use of web services and APIs, following protocols such as CSW, OAI-PMH, ERDDAP, or otherwise, as provided and maintained by the blue data infrastructures. These will be used to deploy machine-to-machine interactions for harvesting metadata, submitting queries, and retrieving resulting metadata, data sets and data products.

The Blue Cloud Data Discovery and Access service should then provide a common interface for discovery and retrieval of data sets and data products from each of the federated blue data infrastructures. The GUI should also include facilities for mapping and viewing the locations of data sets, as this will be part of the query dialogue.

Moreover, conceptually it has been decided to set up the query mechanism as 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 to get more specific data at granule level, using again free search, geographic and temporal criteria, but this time at granule level, and including additional search criteria which are specific per blue data infrastructure
- Finally, users should be 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 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 blue data infrastructures 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 granule level as a second level is applicable to several of the blue data infrastructures, in particular in cases with observation (raw) data which often can be very large collections with numerous data sets. For instance, the SeaDataNet CDI service currently gives discovery and access to more than 2.5 million individual observation data sets for physics, chemistry, geology, biology, geotechnics, and bathymetry. At first level, there are circa 800 CDI aggregated records at collection level, which then give access to the more than 2.5 million granule records, which in the end can be downloaded.

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

In both cases, the overall principle is that queries at level 1 will be applied to search on level 1 over all connected blue data infrastructures and on a common metadata profile, while queries at level 2 will be applied with specific search profiles for each individual blue data infrastructure. For each blue data infrastructure selected records can be included in a shopping basket that then can contain a mix of requested data sets from multiple blue data infrastructures. 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 project 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 blue data infrastructures.

For the data access part of the Blue-Cloud data discovery and access service, 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 blue data infrastructures for level 2 queries, and a shopping mechanism to support the actual discovery and retrieval functions. This part makes use of the experience and software services that MARIS, IFREMER, and EUDAT have developed and are managing for the SeaDataNet CDI service. For the Blue Cloud selected services will be adopted and/or adapted.

Implementing this approach largely depends on the interfaces of blue data infrastructures, that should be supportive. As part of D2.2 the existing web services and APIs of each blue data infrastructure were analysed, tested and documented. In the follow-up further testing and analysis was applied, inter alia for finding the best ways for the deployment of level 2 queries and how to construct the download URLs.

Once launched, the resulting Blue Cloud Data Discovery and Access service will facilitate 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.

It will facilitate managers of blue data infrastructures:

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

This way, the Blue Cloud Data Discovery and Access service will provide a delayed mode service to oversee and to select interesting data sets from the connected blue data infrastructures, followed by downloading and using of the selected data sets by external and VRE users. Next to the offer provided by the blue data infrastructures, the Blue Cloud Data Discovery and Access service, in a later stage of the project, will also index and make available selected data products, resulting from the Blue Cloud demonstrator Virtual Labs, to support a wider distribution and publishing. This will be implemented, once the VRE Products Catalogue has been populated with a critical mass of contents.

2.2 Involved blue data infrastructures

The following blue data infrastructures are pillars under the initial Blue-Cloud developments and services:

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

These blue data infrastructures are mostly complementary to each other, dealing with other data originators and/or different stages in the processing chains from data acquisition to data products to knowledge.

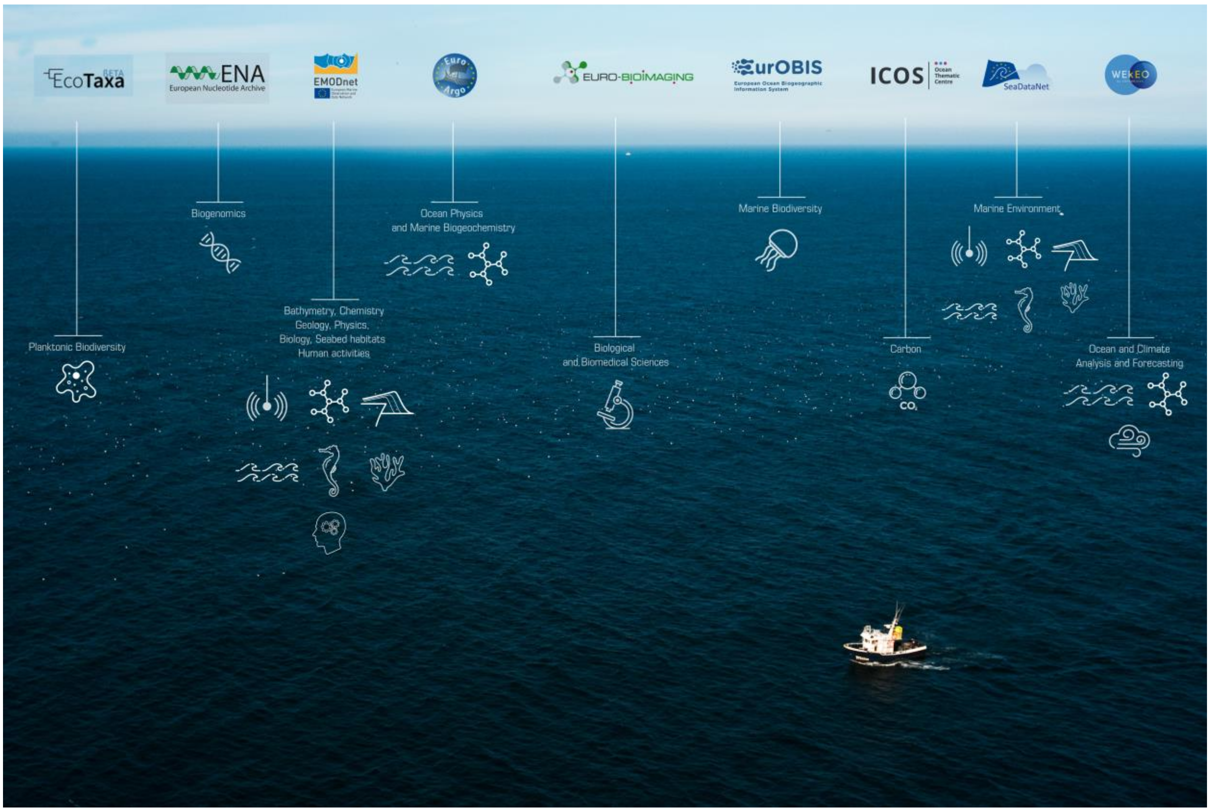


Figure 2.1: Blue data infrastructures as included in initial Blue Cloud discovery and access service

The Blue-Cloud project initially aimed at federating these 10 blue data infrastructures (BDIs). To get a better understanding, each of these existing BDIs infrastructures have been described extensively in deliverable D2.1 - Blue Data Infrastructures – Services Description Report - which was released in February 2020. The report puts a particular focus on their current data discovery and access mechanisms, how these work, types and volumes of data, used formats, which are relevant information as each of the BDIs has to provide endpoints for the federated Blue-Cloud service. From this initial analysis it appeared that the list of BDIs needed to be refined to the following infrastructures: SeaDataNet CDI data service, SeaDataNet Data Products service, EMODnet Chemistry Products service, EurOBIS – EMODnet Biology data collections service, Euro-Argo - Argo GDAC data services, ELIXIR-ENA data service, EuroBioImaging data service, EcoTaxa data service, WekEO (CMEMS) data products service, ICOS-Marine data service and ICOS-SOCAT data service, while EMODnet Bathymetry OGC services will be applied for the global base map in the GUI.

Further analyses activities took place and were reported in Deliverable D2.2 - Blue Data Infrastructures – Services Analysis Report. This focused on their existing data delivery mechanisms and their fitness-for-purpose. As follow-up additional analyses and trials were performed, in particular by MARIS and CNR-IIA, in dialogue with the operators of each of the BDIs. More technical information was gathered and validated concerning the existence and functioning of web services and APIs for detailed queries with additional criteria and at granule level, where relevant, and for actual access to associated data sets. Some BDIs provide multiple web services, sometimes with different results. In those cases, preferred services were selected in communication with their operators. Some use authentication for use of web services and APIs and this could be overcome by registering the Blue-Cloud service as a user, while agreeing to build KPI records of usage as part of the Blue-Cloud shopping component.

From these deeper analyses it appeared that a few of the BDIs cannot be taken onboard of the Blue-Cloud Data Discovery & Access service. This applies for the **EuroBioImaging** service which is being upgraded by EBI and which currently cannot provide the required web services. Hopefully, this BDI can be added later in the remainder of the Blue-Cloud project. It also applies for **WEkEO** which has a well-developed API (so-called HAD). However, WEkEO hosts and gives access to large data products which are resulting from satellite observations, numerical forecast models, and large in-situ collections. These products itself are too big (GBs to TBs) to download as units and are increasing in volume in time. For that reason, the WEkEO API is set-up as a subsetting service, which allows users to compose and run dedicated queries for extracting slices of data from the large products, which then can be downloaded. Emulating this API in the Blue-Cloud discovery service is highly impractical. However as alternative, the WEkEO API is very fit for configuring specific data extraction tasks, for instance in a Jupyter notebook, and this way being included in some of the Blue-Cloud Virtual Labs that require regular input from WEkEO. Therefore, together with MOI, operator of WEkEO, and CNR-ISTI, developer of the Blue-Cloud VRE, it was decided to make a direct coupling of the WEkEO service with the Blue-Cloud VRE using their API embedded in an i-notebook. This is being worked out and tested by CNR-ISTI as part of WP4 developments. Finally, from the analyses it also appeared that ICOS provides two interesting services, namely the ICOS Marine data service and the SOCAT (Surface Ocean CO_2 Atlas) service. It was decided to include both in the Blue-Cloud Data Discovery and Access service.

Following these analyses and observations, the list of BDIs to be incorporated in the first release of the Blue-Cloud Data Discovery & Access service has been confined to:

- SeaDataNet CDI data service (including also CDI entries for EMODnet Chemistry, Bathymetry, and Physics)
- SeaDataNet Data Products service
- EMODnet Chemistry Products service
- 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

So, a total of 9 BDIs for the first release, plus use of EMODnet Bathymetry global map OGC web services for proving a base map in the planned GUI.

2.3 Architecture and modules

The following image gives the architecture of the Blue-Cloud Data Discovery & Access service. The developments are led by MARIS, with contributions of CNR-IIA and EUDAT partners (DKRZ, CINECA, and CSC). MARIS is responsible for the overall system and its integration, while CNR-IIA and EUDAT each deliver components and contributed to the conceptual design, also involving IFREMER and CNR-ISTI. MARIS for a major part makes use of earlier developments and experiences gained during the upgrading of the SeaDataNet CDI service as part of the H2020 SeaDataCloud project. EUDAT develops the Blue-Cloud data delivery component, also using the earlier achievements and experience built up during the SeaDataCloud project.

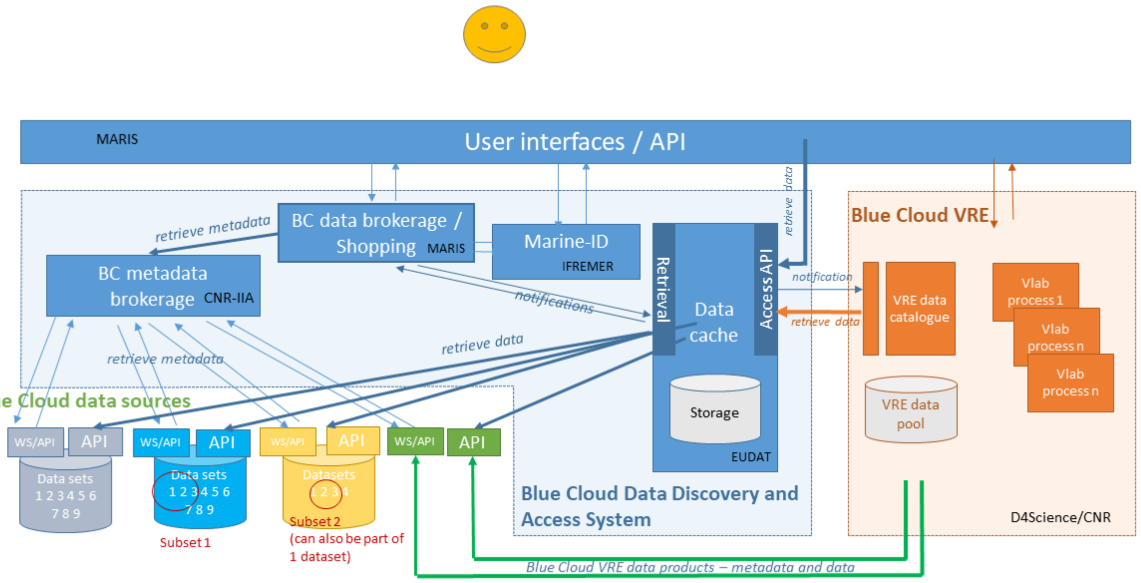


Figure 2.2: Architecture of the Blue Cloud discovery and access service

The Blue Cloud Data Discovery and Access service will consist of a number of modules (services) as indicated in the figure above:

- **Blue Cloud metadata brokerage**, operated by CNR-IIA, dynamically interacting with each of the blue data infrastructures to retrieve, extract and harmonise metadata entries for each blue data infrastructure into a common Blue Cloud level 1 metadata catalogue;
- **Blue Cloud data sources**, comprising blue data infrastructures, that are gathering and managing catalogues and data collections from multiple data and data product originators; the Blue-Cloud VRE is also considered as Blue Cloud data source, concerning publishable data products as resulting from the demonstrator Virtual Labs.
- **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 from the Blue Cloud data brokerage, and following the instructions as received from 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, 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 blue data source to allow the first level queries at collection level with a few selection criteria and this way to identify which of the blue data infrastructures 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 blue data infrastructures that only have data collections and other data products, but no service at granule level. While for other blue data infrastructures, supporting deeper searching at level 2 – granule level -, customised search profiles have been formulated, which allow the data broker to interacvt with the provided web services and APIs of the blue data infrastructures. The Blue Cloud Data brokerage service also contains a shopping mechanism with basket and ledger, by which users (external users and VRE) and blue data infrastructures can be informed about shopping transactions and their status in time. It interacts with the Blue Cloud Data Cache to give it precise instructions about retrieving data sets from the blue data infrastructures 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 and shopping transactions at level 2, and to provide access to the transaction ledger for users and blue data infrastructures.

Note: In figure 2.2, the Blue Cloud VRE is given with simplified graphics, only to underpin the exchange with the Blue Cloud Data Discovery and Access service. Much more detail about the Blue Cloud VRE will be given in the next chapter.

3 Level 1 – Common Blue-Cloud metadata catalogue at collection level

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 DAB technology. This middleware harvests metadata at collection level from each of the blue data infrastructures (BDIs), using their indicated web services or APIs. The DAB service then 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 CSW service with a common XML profile for each BDI. See image below.

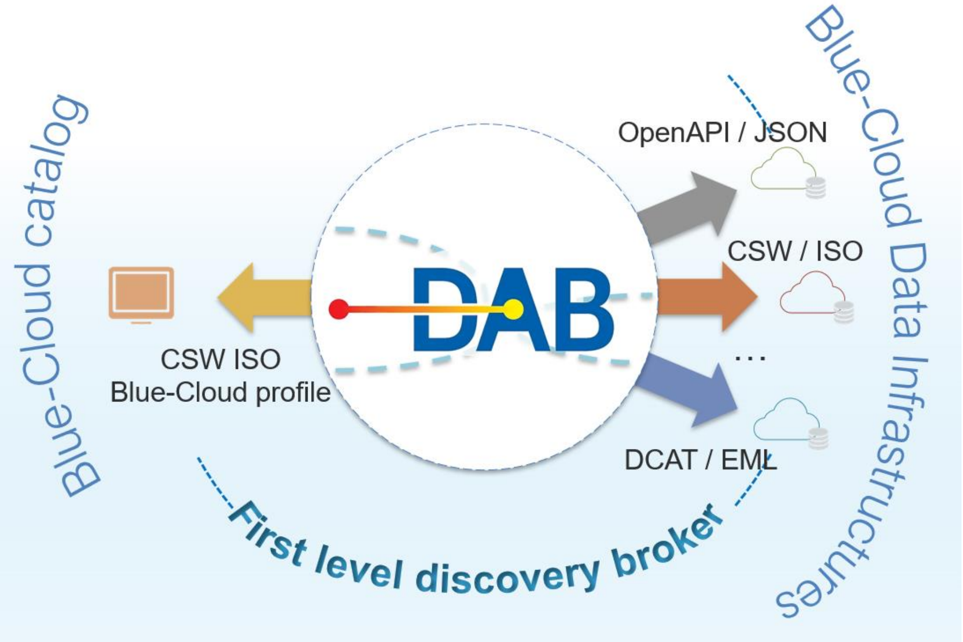


Figure 3.1: Blue-Cloud first level discovery broker component harmonizes the protocols and data models published by different heterogeneous BDIs to a harmonized CSW service based on ISO 19115.

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

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

- INSTRUMENT: measuring instrument used to acquire the data
- PLATFORM: platform from which the data were taken
- ORGANIZATION: organization associated with the collection
- DATESTAMP: the latest update date of the metadata description
- REVISION_DATE: the latest update date of the data
- RESOURCE_LINKS: download links where available and useful

The DAB service of CNR-IIA regularly harvests and thus updates the output of the Blue-Cloud CSW ISO v. 2.0.2 service per BDI. CSW is a well-known standard web service of the Open Geospatial Consortium (OGC), recommended by many initiatives for sharing metadata on the web**.** Next, 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 then 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 DBIs 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 Discovery & Access service. 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, namely the SeaDataNet data products, the EMODnet Chemistry data products, and the EurOBIS – EMODnet Biology data collections. All has been set up as an automatic process without human intervention, driving daily 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 Data Discovery and Access service.

In the next section the implementation of the harmonized discovery interface based on CSW ISO service and Blue-Cloud ISO 19115 profile is presented. The other following sections, one for each BDI, describes the implementation of the components needed to harvest and mediate data from each different BDI.

3.1 CSW for ISO Blue-Cloud profiles per blue data infrastructure

The DAB CSW provides internal endpoints with the harmonised Blue-Cloud metadata profile at collection level for each of the blue data infrastructures. These are meant as internal services as MARIS harvests from these DAB CSW services for extra processing, such as filtering out some records without reliable data links, modifying selected data links for bypassing web forms, and others.

- DAB Service endpoint (global): https://blue-cloud.geodab.eu/gs-service/services/essi/view/blue-cloud/csw
  - Service endpoint (Euro Argo
  - - Argo): https://blue-cloud.geodab.eu/gs-service/services/essi/view/argo/csw
  - Service endpoint (ELIXIR-ENA): https://blue-cloud.geodab.eu/gs-service/services/essi/view/elixir-ena/csw
  - Service endpoint (EMODnet Chemistry): https://blue-cloud.geodab.eu/gs-service/services/essi/view/emodnet-chemistry/csw
  - Service endpoint (EurOBIS): https://blue-cloud.geodab.eu/gs-service/services/essi/view/eurobis/csw

  - Service endpoint (ICOS Data Portal): https://blue-cloud.geodab.eu/gs-service/services/essi/view/icos-data-portal/csw
  - Service endpoint (ICOS SOCAT): https://blue-cloud.geodab.eu/gs-service/services/essi/view/icos-socat/csw
  - Service endpoint (SeaDataNet CDI service): https://blue-cloud.geodab.eu/gs-service/services/essi/view/seadatanet-open/csw
  - Service endpoint (SeaDataNet Products): https://blue-cloud.geodab.eu/gs-service/services/essi/view/seadatanet-products/csw
- Service interface: CSW ISO v. 2.0.2
- Data model: Blue-Cloud ISO 19115 profile

The service publishes OGC CSW service catalogue standard, a well-known standard service recommended by many initiatives for sharing metadata on the web.

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

The relevant Blue-Cloud metadata can be found by MARIS in the returned records at the XPath locations reported in the following table.

Metadata element	Path
IDENTIFIER	//gmd:fileIdentifier/gco:CharacterString
TITLE	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:citation/gmd:CI_Citation/gmd:title/*[1]
ABSTRACT	//gmd:abstract
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]
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
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_DataIdentificati

	on/gmd:extent/gmd:EX_Extent/gmd:temporalElement/gmd:EX_TemporalExtent/gmd:extent/gml32:TimePeriod/gml32:endPosition
PARAMETER	/gmi2019:MI_Metadata/gmd:contentInfo/gmi2019:MI_CoverageDescription/gmd:attributeDescription/gco:RecordType
INSTRUMENT	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords[gmd:type/gmd:MD_KeywordTypeCode/@codeListValue='instrument']/gmd:keyword/*[1]
PLATFORM	/gmi2019:MI_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords[gmd:type/gmd:MD_KeywordTypeCode/@codeListValue='platform']/gmd:keyword/*[1]
ORGANIZATION	//gmd:CI_ResponsibleParty/gmd:organisationName/*[1]
DATESTAMP	/gmi2019:MI_Metadata/gmd:dateStamp/gco:Date
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

Although the core Blue-Cloud metadata element set is composed by 12 elements only, the returned metadata records optionally can contain many other elements as described in ISO 19115 parts 1 and 2, that are useful to describe the collections in full details. These however could be not present for all the BDIs and it’s not assured that each BDI uses these elements in the same way (e.g. metadata elements might be encoded differently).

The following paragraphs give the details per BDI which services were used by the DAB service for harvesting the original metadata at collection level.

3.2 Euro Argo – ARGO GDAC service

The EuroArgo ERIC allows active coordination and strengthening of the European contribution to the international Argo program. Its main objectives are to provide, deploy and operate the European contribution to the global array of Argo floats (currently around 800 floats, ¼ of the global array) and an enhanced coverage of European seas, to expand towards biogeochemistry, greater depths and high latitudes and to provide access to quality-controlled data and derived products. The EuroArgo ERIC also provides access to quality-controlled data and derived products. This is done by Ifremer in France by hosting one of the two Global Data Assembly Centres (GDACs). The Argo GDACs assemble all data observed by the global array and distribute them worldwide both in real time and in delayed mode. This implicates that the EuroArgo GDAC at Ifremer gives access to the global Argo dataset. This currently concerns a few million data sets from about 16.000+ Argo floats.

- Service endpoint: https://fleetmonitoring.euro-argo.eu/
- Service interface: Swagger 2.0
- Data model: custom, JSON based

A JSON based API is used to discover metadata about the ARGO floats. 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.

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
	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

3.3 ELIXIR-ENA service

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.

- Service endpoint: https://www.ebi.ac.uk/ena/portal/api/
- Service interface: Swagger 2.0
- Data model: custom, JSON based

A Swagger based API is published by EMBL-EBI to discover metadata about the available studies. In this case ELIXIR-ENA “study” level records are mapped to the Blue-Cloud first level (collection) metadata records. A set of predefined studies and study collections are selected to be harvested. Currently they are 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 contains the latitude and longitude of the specific acquisition), to complement study-level metadata.

In particular, 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.

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

3.4 EMODnet Chemistry Data Products service

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

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

The products are described with metadata in the Chemistry products catalogue, have DOIs and landing pages for citation, and can be downloaded.

- Service endpoint: https://sextant.ifremer.fr/geonetwork/srv/eng/csw-EMODNET_Chemistry
- Service interface: CSW ISO v. 2.0.2
- Data model: ISO 19115 based profile

EMODnet Chemistry exposes standard CSW ISO interface that 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

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	Seems to be not available for this BDI
Platform	Seems to be 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

3.5 EurOBIS – EMODnet Biology service

EurOBIS was developed by the Flanders Marine Institute in 2004, within the framework of the MarBEF project (MARine Biodiversity and Ecosystem Functioning). It brings together biogeographic data collected within European marine waters, or by European researchers and institutes outside Europe. It focuses on taxonomy and distribution records in space and time and offers a number of online tools to easily query and visualise the data. Currently, EurOBIS holds 850+ 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 Information System (OBIS) and eventually become available via the Global Biodiversity Information Facility (GBIF), hosting global marine and terrestrial distribution data. From 2009 EurOBIS became the backbone of the European Marine Observation and Data Network Biology (EMODnet Biology), allowing a flow of EurOBIS data through its portal. In 2014 EurOBIS became part of the central Species Information Backbone of LifeWatch, which aims at standardizing species data and integrating the distributed biodiversity data and taxonomic repositories and operating facilities as well as filling the gaps in our knowledge. The EurOBIS data management team is supported by LifeWatch Belgium, part of the European LifeWatch E-Science Infrastructure for Biodiversity and Ecosystem Research.

- Service endpoint: http://www.vliz.be/imis
- Service interface: DCAT
- Data model: RDF Turtle; EML 2.1.1

The available datasets published by the DCAT service can be obtained executing the following request:

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

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

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

Each EML document is then 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	EurOBIS metadata element
Identifier	/eml/dataset/alternateIdentifier
Title	/eml/dataset/title

Keyword	/eml/dataset/keywordSet/keyword
Bounding box	/eml/dataset/coverage/geographicCoverage/boundingCoordinates/westBoundingCoordinate
	/eml/dataset/coverage/geographicCoverage/boundingCoordinates/eastBoundingCoordinate
	/eml/dataset/coverage/geographicCoverage/boundingCoordinates/northBoundingCoordinate
	/eml/dataset/coverage/geographicCoverage/boundingCoordinates/southBoundingCoordinate
Temporal extent	/eml/dataset/coverage/temporalCoverage/rangeOfDates/beginDate/calendarDate
	/eml/dataset/coverage/temporalCoverage/rangeOfDates/endDate/calendarDate
Parameter	Seems to be not available for this BDI
Instrument	Seems to be not available for this BDI
Platform	Seems to be not available for this BDI
Organization	/eml/dataset/creator/organizationName
Date stamp	//*:dateStamp
	Temporal extent end
Revision date	Temporal extent end

3.6 ICOS Marine Data Portal service

ICOS is an international organisation of thirteen European member countries and over 130 greenhouse gas measurement stations aimed at quantifying and understanding the greenhouse gas balance of Europe and neighbouring regions. ICOS data is made available at the Carbon Portal, a one-stop shop for all ICOS data products. The Ocean Thematic Centre is one of four central facilities within the European research infrastructure Integrated Carbon Observation System (ICOS). The marine element of ICOS provides long-term oceanic observations, which are required to understand the present state and better predict future behaviour of the global carbon cycle and climaterelevant 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. The ICOS Carbon Portal provides observation data from over 130 greenhouse gas measurement stations.

- Service endpoint: https://meta.icos-cp.eu/sparqlclient/
- Service interface: SPARQL endpoint
- Data model: custom, JSON based

A SPARQL endpoint base API is used to retrieve the list of metadata “cruise” files. An HTTP-POST request is executed with the following filters for marine observation data:

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

The list of available information coming from the previous request is used to build the HTTP-GET request to retrieve full metadata for each "cruise”. Example given to obtain the metadata record for cruise with fileName *58G2_NRT_20210118.csv* the following HTTP-GET request is executed:

https://meta.icos-cp.eu/objects/1Mr0pKFtpDMP8pPymKfXMnCQ/58G2_NRT_20210118.csv.json

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

Example mappings for some common metadata fields from the JSON response:

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’]
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’]

3.7 ICOS SOCAT service

SOCAT stands for Surface Ocean CO_2 Atlas. The SOCAT data set serves a wide range of user communities. The SOCAT version 2020, made public in 2020, includes data from more than 10 countries. It has 28.2 million quality-controlled surface ocean fCO_2 measurements from 1957 to 2020. The SOCAT data set uses IOCCP recommended formats for metadata and data reporting. SOCAT quality control is carried out by regional working groups with a global group for coordination. Through the SOCAT portal there are around 6000 trajectories available from 1957-2018 with 26 million data values.

- Service endpoint: https://ferret.pmel.noaa.gov/socat/erddap/tabledap/socat_v2020_fulldata
- Service interface: ERDDAP
- Data model: custom, XHTML based

An XML table of all the ICOS SOCAT metadata records can be retrieved with the following call:

https://data.pmel.noaa.gov/socat/erddap/tabledap/socat_v2020_fulldata.xhtml?expocode%2Cdataset_name%2Cplatform_name%2Cplatform_type%2Corganization%2Cgeospatial_lon_min%2Cgeospatial_lon_max%2Cgeospatial_lat_min%2Cgeospatial_lat_max%2Ctime_coverage_start%2Ctime_coverage_end%2Cinvestigators%2Csocat_doi&distinct()

Each row of the above table 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	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

Organization	organization
Date stamp	time_coverage_end
Revision date	time_coverage_end

3.8 SeaDataNet CDI service

SeaDataNet 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, 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 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. A core SeaDataNet service is the **Common Data Index (CDI) data discovery and access service** which provides harmonized discovery and access to a large volume of marine and ocean data sets, both from research and monitoring organisations, which increasingly are major input for developing added-value services and products that serve users from government, research and industry. 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 **2.5 Million data sets**, originating from more than **850 organisations** in Europe, covering physical, geological, chemical, biological and geophysical data, and acquired in European waters and global oceans. For the Blue-Cloud access will be given to the open data (ca 85% of all SeaDataNet CDI data) and starting at a collection level.

- Service endpoint: https://cdi.seadatanet.org/report/aggregation/open
- Service interface: CDI XML inventory
- Data model: SeaDataNet CDI ISO 19115 profile

The service provides an inventory of CDI XML documents, structured as a XML encoded list of URLs. It is possible to harvest all the records simply issuing HTTP-GET requests of each individual URLs. Example given:

https://cdi.seadatanet.org/report/aggregation/486/486/486/4/ds03/open/xml

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 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_Geographic BoundingBox
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
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

3.9 SeaDataNet Data Products service

SeaDataNet also generates several derived data products such as tens of aggregated data collections and climatologies, such as for Temperature & Salinity. These are provided through a catalogue service for SeaDataNet data products.

- Service endpoint: https://sextant.ifremer.fr/geonetwork/srv/eng/csw-SEADATANET
- Service interface: CSW ISO v. 2.0.2
- Data model: ISO 19115 based profile

SeaDataNet Products exposes standard CSW ISO interface that 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	Seems to be not available for this BDI
Platform	Seems to be 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_Date[gmd:dateType/gmd:CI_DateTypeCode/@codeListValue='revision']/gmd:date

3.10 EcoTaxa 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, high-throughput imaging of plankton and of the Oceanomics project which covered the exploitation of data collected during the Tara Oceans cruise, including quantitative imaging. It is now developed mainly through the WWWPIC project funded by the Belmont Forum and as part of the Blue-Cloud project. The aim of EcoTaxa is to centralize images of plankton, to allow their collaborative sorting along a universal taxonomy and to accelerate it through machine learning. It produces ecological data in the form of concentration and biovolume of organisms in a given taxon, at a given station (lat, lon, time). Visitors have free access to the specimens that have been already identified by taxonomist experts. They can explore the database by navigating along the UniEuk taxonomic tree which aims at unifying taxonomic names and tree according to reliable and curated molecular phylogenies. It encompasses the whole Eukaryotic and Prokaryotic lineages (Viruses coming soon) that have been molecularly described. Then images can be filtered according to several sample criteria. Tools are provided to support the annotation of large image datasets by supervised machine learning prediction. Currently, EcoTaxa hosts 130 million objects, 55 millions of which have been classified by a human operator. Those objects are organised in ~2000 projects, have been uploaded from ~300 organisations and classified by ~900 users. Those objects are contained in samples that come from all over the world.

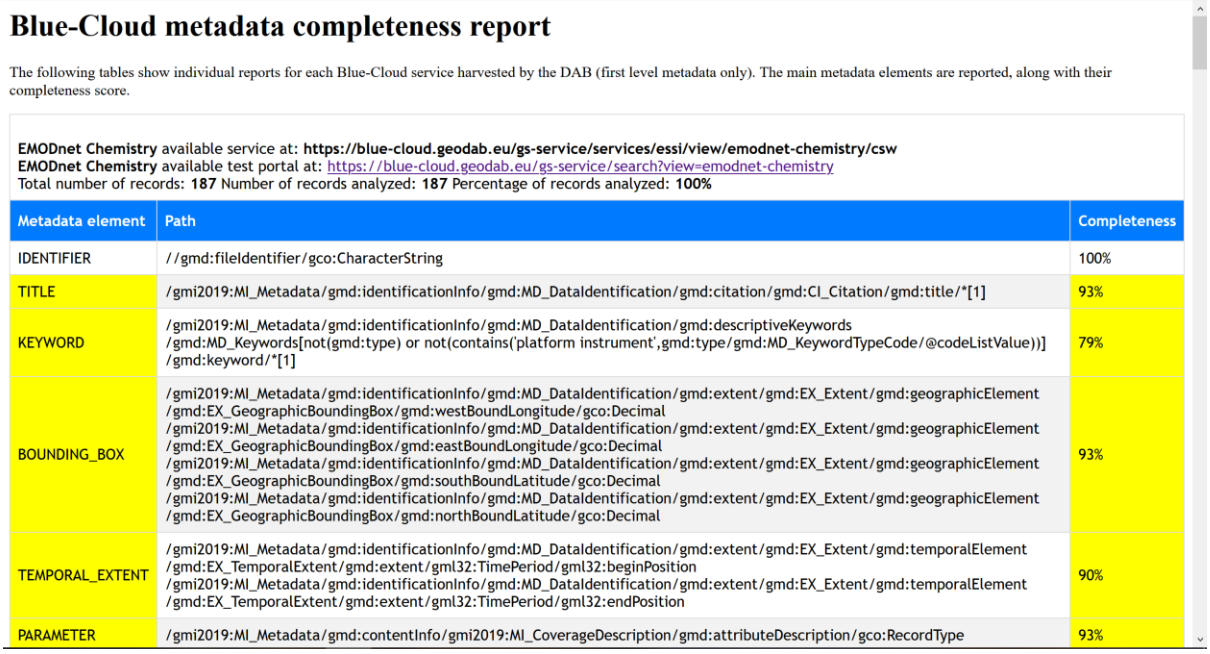
It was decided to provide *summarised data* from EcoTaxa (concentrations per sample = lat/lon/time point) to EurOBIS, so that the Blue-Cloud can fetch the metadata about those datasets from EurOBIS. Blue-Cloud users should be able to discover the EcoTaxa datasets through EurOBIS and then, if more details are needed, turn to EcoTaxa to get them at granule (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 the first level of EcoTaxa in the Blue-Cloud service, the subset of 10 records is filtered from the full EurOBIS collection records. See paragraph 3.5 for the EurOBIS mapping.

3.11 Blue-Cloud first level (collection) metadata report – internal service

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



Blue-Cloud metadata completeness report

The following tables show individual reports for each Blue-Cloud service harvested by the DAB (first level metadata only). The main metadata elements are reported, along with their completeness score.

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: **187** Number of records analyzed: **187** 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]	93%
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]	79%
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	93%
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	90%
PARAMETER	/gmi2019:MI_Metadata/gmd:contentInfo/gmi2019:MI_CoverageDescription/gmd:attributeDescription/gco:RecordType	93%

Figure 3.2: DAB completeness report of core metadata elements in each of the blue data infrastructures as determined from their web services or APIs

The report shows with graphical indicators the most recent status of the Blue-Cloud metadata content. There are listed 8 different sections: one for each BDI. For each one is reported the total number of collection records available and the presence of each of the 12 Blue-Cloud core metadata, highlighting in yellow and red possibly warnings and errors due to a too low occurrence percentage. These are the elements on which operators of blue data infrastructures should focus to improve their service metadata quality.

Currently the summary results are as shown by the following table.

BDI	Total records
Euro Argo – Argo GDAC	15997
ELIXIR-ENA	28
EMODnet Chemistry	187
EurOBIS – EMODnet Biology	862
ICOS Data Portal	130
ICOS SOCAT	6471

SeaDataNet CDI	795
SeaDataNet Products	50
EcoTaxa collections	10

Table: Numbers of data collection records as harvested by the DAB service from each of the blue data infrastructures through their web services or APIs

4. Level 2 – Individual search profiles at granule level

Following the technical documentation, collated in D2.2, and further analyses and contacts with the operators of the blue data infrastructures, MARIS has progressed with integrating each BDI as part of the data brokerage service, arranging the 2nd query level and direct download links. The preferred way forward is that use can be made 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. And again, all these functions in automated and dynamic machine-to-machine interactions between the Blue-Cloud Data Discovery & Access service and the web services or APIs of each BDI.

In practice, deploying this preferred concept is not feasible for each BDI as their web services are not (yet) fit and could not be adapted on short-term. In those cases, alternatives are followed, 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 implicates that for each BDI a customised plug-in is being developed and deployed, also taking into account performance. In addition, hurdles had to be 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, which altogether took more time than earlier envisaged.

The following table gives an overview of the web services or APIs of each BDI for which the data broker service has set up customized interfaces.

Blue Data Infrastructure	**Type of web service / API for level 2**	**Remarks**
SeaDataNet CDI service	Dedicated API (Swagger documented)	Granule level
SeaDataNet Data Products service	DAB CSW	Using Blue-Cloud common level 1 catalogue + solution to by-pass web forms in download links
EMODnet Chemistry Products service	DAB CSW	Using Blue-Cloud common level 1 catalogue + solution to by-pass web forms in download links
EurOBIS – EMODnet Biology data collections service	DAB CSW	Using Blue-Cloud common level 1 catalogue + solution to construct download

		links from additional web services
ELIXIR-ENA data service	Dedicated API	Granule level
EcoTaxa data service*	Dedicated API (OpenAPI)	Granule level
ICOS-Marine data service	SPARQL service / ERDDAP	Granule level
ICOS-SOCAT data service*	ERDDAP	Granule level
Euro-Argo - Argo GDAC data services	Dedicated API	Solution to find FTP folders for download links. Granule level

The services, marked with * are under development; however, it is expected that these will be delivered soon too.

4.1 Euro Argo – ARGO GDAC service

The Argo GDAC at EuroArgo ERIC gives access to metadata and data sets from circa 16.000 Argo floats (= collection level) and the data sets from their cycles (= granule level), which comprise several millions of data sets and often multiple data files per cycle.

- Service endpoint: https://fleetmonitoring.euro-argo.eu/
- Service interface: Swagger 2.0
- Data model: custom, JSON based

The JSON based API is used to discover metadata about the ARGO floats and their cycles with additional metadata. It is possible to harvest metadata for all the floats and all their cycles by extracting the JSON records. These are dynamically stored by MARIS in an SQL database and then indexed to an Elastic Search database, creating several facets.

Example of full metadata per Float ID:

https://fleetmonitoring.euro-argo.eu/floats/6903238
with 6903238 is a unique Float ID.

The associated data files can be found in an FTP service:

- ftp://ftp.ifremer.fr/ifremer/argo

This allows to locate a folder for each float and cycle combination and then to download the related data files.

Based upon the harvested metadata, search facets can be generated at granule level and searching can be done by:

- Free search
- Lat-Lon box

- Date period
- Platform type
- Country
- Status Code
- Deployment year
- Transmissionsystem
- Datacenter Name
- Variables
- Networks
- Owner
- Projects

4.2 ELIXIR-ENA service

A Swagger based API is published by EMBL-EBI to discover metadata about the available studies.

- Service endpoint: https://www.ebi.ac.uk/ena/portal/api/
- Service interface: Swagger 2.0
- Data model: custom, JSON based

In this case ELIXIR-ENA “study” level records are mapped to the Blue-Cloud first level (collection) metadata records. A set of predefined studies and study collections are selected to be harvested. Currently they are the following ones:

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

For each study, it is possible to query for related data sets, using the API and locating the dsata links to these study results. In this case, there is direct interaction with the API for each action at the 2nd level. The query criteria are limited to:

- Free search
- Lat-Lon box
- Date period

4.3 EMODnet Chemistry Data Products service

For the EMODnet Chemistry products, there are no granules, as users will download complete data products. The second level queries are enriched with additional criteria from the common Blue-Cloud collection metadata model to make it easier to find the required data products. Use is made again of the level 1 metadata catalogue database that MARIS has created on top of the DAB CSW services. Using Elastic Search, facets are generated and included in the 2nd level Search interface. The searching can be done by:

- Free search
- Lat-Lon box
- Date period

- Keywords
- Parameters
- Organisation

The data links are provided by the DAB derived catalogue; however, these URLs need to be modified in order to by-pass the call up of a web form. This by-pass has been arranged with the operator of the blue data infrastructure.

4.4 EurOBIS – EMODnet Biology service

For the EurOBIS – EMODnet Biology service, there are no granules, as users will download complete data products. The second level queries are enriched with additional criteria from the common Blue-Cloud collection metadata model to make it easier to find the required data products. Use is made again of the level 1 metadata catalogue database that MARIS has created on top of the DAB CSW services. Using Elastic Search, facets are generated and included in the 2nd level Search interface. The searching can be done by:

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

The data links are provided by the DAB derived catalogue; however, to find these URLs some extra activities were required to construct the EurOBIS download links to Validated Data Records. Use is made of:

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

which gives metadata on all EurOBIS resources, with links like:

http://ipt.vliz.be/eurobis/resource?r=dasid_6512#Dataset

http://ipt.vliz.be/eurobis/resource?r=benthos_azov_sea#Dataset

Then, one can ask for the related EML file by replacing 'resource?r' by '**eml.do?r**' like:

http://ipt.vliz.be/eurobis/eml.do?r=dasid_6512#Dataset

http://ipt.vliz.be/eurobis/eml.do?r=benthos_azov_sea#Dataset

In each EML file, there is a Download link in the following section, like:

https://www.emodnet-biology.eu/portal/index.php?dasid=4755 which leads to the unique and persistent DASID of each EurOBIS record.

With this DASID one can compose the direct download link to the validated data sets by means of an OGC WFS link:

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+(4755);context:0100&outputFormat=csv

This script has been deployed by the DAB service, so that MARIS can retrieve the correct data links directly from the DAB service output.

4.5 ICOS Marine Data Portal service

For the level 2 queries, use should be made of the SPARQ endpoint.

- Service endpoint: https://meta.icos-cp.eu/sparql
- Service interface: SPARQL endpoint
- Data model: custom, JSON based

Alternatively, ICOS is developing an ERDDAP endpoint for the ICOS data sets.

Currently, the analysis by MARIS is underway for determining the best way forward for accessing the ICOS Marine data service at the granule level. This will be solved in the near future.

4.6 ICOS SOCAT service

The SOCAT service is accessible by an ERDDAP deployment.

- Service endpoint: https://ferret.pmel.noaa.gov/socat/erddap/tabledap/socat_v2020_fulldata
- Service interface: ERDDAP
- Data model: custom, XHTML based

ERDDAP can be easily used as a middle ware to retrieve the full metadata and build facets, using again Elastic Search. This has been implemented by MARIS to enable granule searching by:

- Free search
- Lat-Lon box
- Date period
- Platform Type
- Platform Name
- Organization
- Qc Flag
- Investigators

The data links are also provided by the ERDDAP service as REST URLs.

4.7 SeaDataNet CDI service

A dedicated API is under development for SeaDataNet CDI service at granule level, in particular for open data, and this works in combination with the CDI shopping mechanism. Therefore, Blue-Cloud is seen as a CDI user with a Marine-ID. This API is still under development, in particular, its documentation in Swagger. However, it could already be used in the Blue-Cloud as MARIS is also the developer of the CDI service API.

- Service endpoint: https://cdi-open.seadatanet.org/api
- Service interface: Swagger 2.0

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 granule searching by:

- Free search
- Lat-Lon box
- Date period
- Measuring area type
- Discovery Parameter (P02)
- Parameter Group (P03)
- Discipline (P08)
- Organization

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.

4.8 SeaDataNet Data Products service

For the SeaDataNet data products, there are no granules, as users will download complete data products. The second level queries are enriched with additional criteria from the common Blue-Cloud collection metadata model to make it easier to find the required data products. Use is made again of the level 1 metadata catalogue database that MARIS has created on top of the DAB CSW services. Using Elastic Search, facets are generated and included in the 2nd level Search interface. The searching can be done by:

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

The data links are provided by the DAB derived catalogue; however, these URLs need to be modified in order to by-pass the call up of a web form. This by-pass has been arranged with the operator of the blue data infrastructure.

4.9 EcoTaxa service

EurOBIS is publishing EcoTaxa “Collections”. Therefore, it was decided that an EcoTaxa Application Programming Interface (API) should be developed as part of the Blue-Cloud project to allow for browsing and extracting EcoTaxa information at granule level. This way, Blue-Cloud users should be able to discover the EcoTaxa datasets through EurOBIS and then, if more details are needed, turn to the 2nd level search, driven by the EcoTaxa API to get them at granule (i.e. sample or object) level.

The specifications of the EcoTaxa API have been documented in D2.3 - EcoTaxa data service, which was released in M15. In the meantime, good progress has been made with its development.

- Service endpoint documentation: https://ecotaxa.obs-vlfr.fr/api/docs
- Service interface: OpenAPI/Swagger

Its specification is open and it benefits from many functionalities from this widely used toolbox. The interested user will browse the documentation URL to find the latest API endpoints and the parameters they require.

Using the API requires authentication and for that purpose a Blue-Cloud user account was arranged with the operator of EcoTaxa, giving access to the EcoTaxa projects and objects that are related to the EcoTaxa collection records in EurOBIS.

Currently, the analysis by MARIS is underway for determining the best way forward for accessing and using the EcoTaxa API at granule level. This will be solved in the near future.

5. Delivery service

The Blue-Cloud Data Discovery and Access service works with a shopping mechanism. This has been adopted and adapted from 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; to be expanded with functionality for BDI operators to check Blue-Cloud orders made for their BDIs.
- E-mails to users to confirm their submitted orders and to alert users when orders are completed and ready for downloading.

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 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 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. The communication between Flask APIs and the Celery tasks is ensured by a message broken system based on RabbitMQ.

The plan is that retrieved data will be accessed by two different paths. Next to the direct download by users (see description above), it is also planned 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 will take 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 planned workflow will be:

- 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 notifies the VRE that a data batch is available for pull (asynchronous request from EUDAT to VRE API, including download link and user name).
- The VRE downloads the data using the specified link (request from VRE to EUDAT API).
- Then, the data can be deleted or kept for further use (e.g. later downloads or transfers). This will depend on policies as well as on technical considerations and is not decided yet.

To realize this workflow, a component will have to be developed and added by CNR-ISTI to the VRE that can handle requests to pull data. As both CNR-ISTI and EUDAT’s participating data centres DKRZ, CINECA and CSC are part of the European research network GÉANT, this transfer will benefit from the very fast network backbone between major European research centres. Note: this functionality has not yet been deployed, but this development will be undertaken soon after the launch of the Blue-Cloud Data Discovery and Access service.

6. Deployment of the Blue-Cloud Data Discovery and Access service

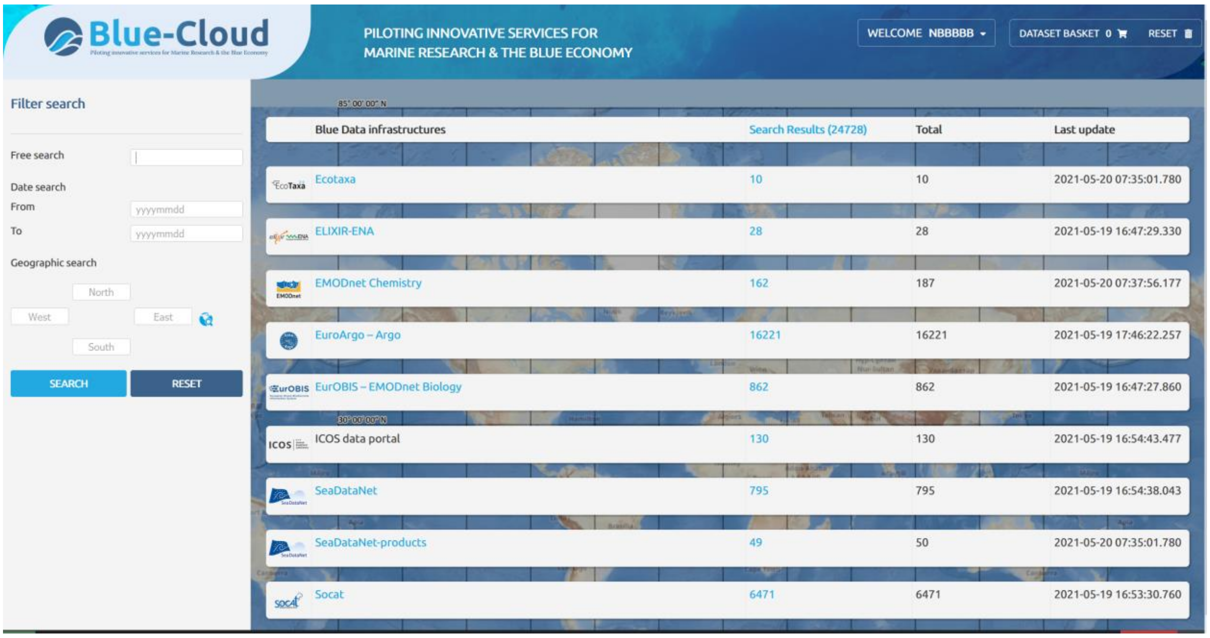
The development of the Beta-version of the Blue-Cloud Data Discovery and Access service is making very good progress and it is planned to launch it as operational service in the middle of June 2021. This will give some extra time for solving the integration at level 2 of the two missing blue data infrastructures, ICOS and EcoTaxa, finalising some open items, and testing the integrated system.

A landing page will be implemented at:
https://data.blue-cloud.org

The landing page will give an introduction and short explanation about the working of the service as well as the option for registering as Marine-ID user, which is required for submitting data requests and for having access to the MyBlueCloud dashboard.

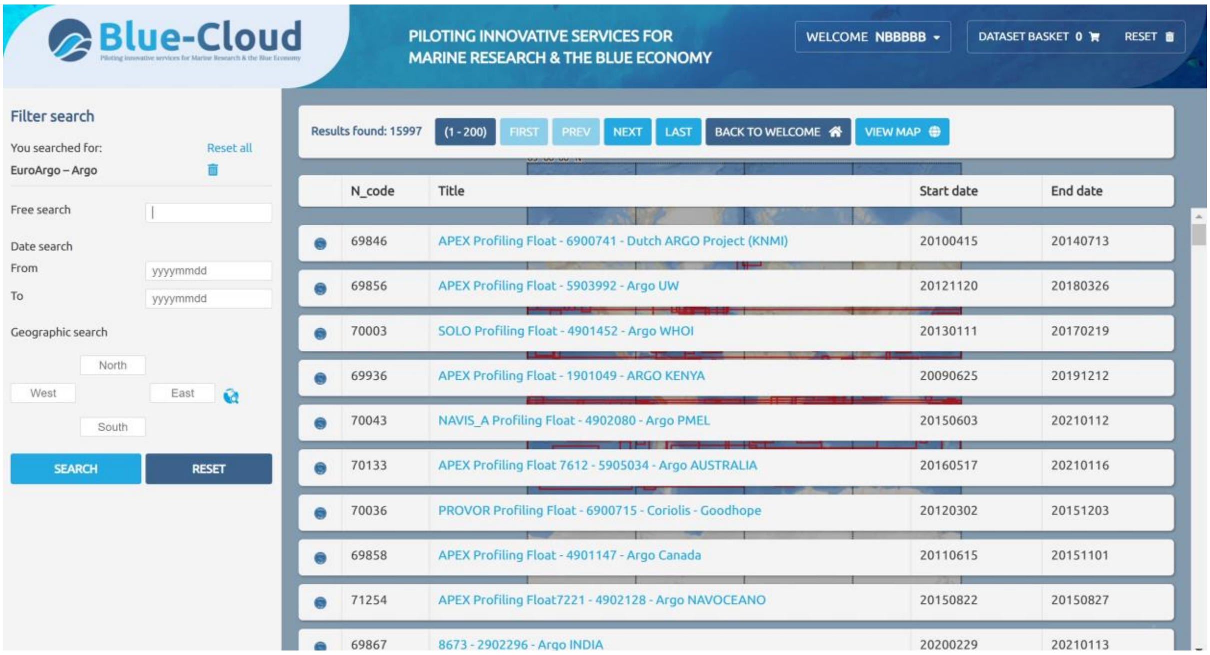
The Blue-Cloud Data Discovery and Access service will be made available at: https://data.blue-cloud.org/search

The following images give some screengrabs from the Blue-Cloud Data Discover and Access service under development.



Blue Data infrastructures	Search Results (24728)	Total	Last update
Ecotaxa	10	10	2021-05-20 07:35:01.780
ELIXIR-ENA	28	28	2021-05-19 16:47:29.330
EMODnet Chemistry	162	187	2021-05-20 07:37:56.177
EuroArgo – Argo	16221	16221	2021-05-19 17:46:22.257
EurOBIS – EMODnet Biology	862	862	2021-05-19 16:47:27.860
ICOS data portal	130	130	2021-05-19 16:54:43.477
SeaDataNet	795	795	2021-05-19 16:54:38.043
SeaDataNet-products	49	50	2021-05-20 07:35:01.780
Socat	6471	6471	2021-05-19 16:53:30.760

Figure 6.1: Home page of the service with overview of connected Blue Data Infrastructures and options for level 1 (collections) search


Blue-Cloud
Piloting innovative services for Marine Research & the Blue Economy
PILOTING INNOVATIVE SERVICES FOR
MARINE RESEARCH & THE BLUE ECONOMY
WELCOME NBBBBB
DATASET BASKET 0
RESET
SEARCH
RESET
(1 - 200)
FIRST
PREV
NEXT
LAST
BACK TO WELCOME
VIEW MAP

Filter search

You searched for: EuroArgo – Argo

Reset all

Free search

Date search

From yyyymmdd

To yyyymmdd

Geographic search

North

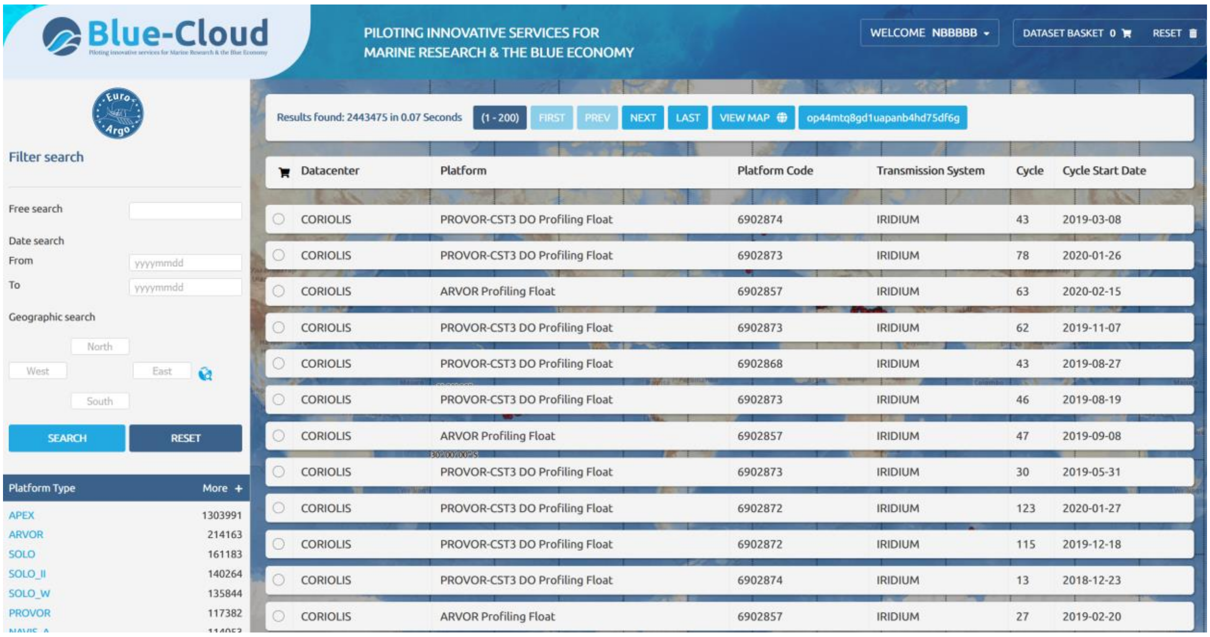
West East

South

Results found: 15997

	N_code	Title	Start date	End date
●	69846	APEX Profiling Float - 6900741 - Dutch ARGO Project (KNMI)	20100415	20140713
●	69856	APEX Profiling Float - 5903992 - Argo UW	20121120	20180326
●	70003	SOLO Profiling Float - 4901452 - Argo WHOI	20130111	20170219
●	69936	APEX Profiling Float - 1901049 - ARGO KENYA	20090625	20191212
●	70043	NAVIS_A Profiling Float - 4902080 - Argo PMEL	20150603	20210112
●	70133	APEX Profiling Float 7612 - 5905034 - Argo AUSTRALIA	20160517	20210116
●	70036	PROVOR Profiling Float - 6900715 - Coriolis - Goodhope	20120302	20151203
●	69858	APEX Profiling Float - 4901147 - Argo Canada	20110615	20151101
●	71254	APEX Profiling Float7221 - 4902128 - Argo NAVOCEANO	20150822	20150827
●	69867	8673 - 2902296 - Argo INDIA	20200229	20210113

Figure 6.2: Browse of Argo records at level 1 with options for retrieving metadata details of each collection and queries at level 1


Blue-Cloud
Piloting innovative services for Marine Research & the Blue Economy
PILOTING INNOVATIVE SERVICES FOR
MARINE RESEARCH & THE BLUE ECONOMY
WELCOME NBBBBB
DATASET BASKET 0
RESET
Euro Argo
SEARCH
RESET
(1 - 200)
FIRST
PREV
NEXT
LAST
VIEW MAP
op44mtq8gd1uapanb4hd75df6g

Filter search

Free search

Date search

From yyyymmdd

To yyyymmdd

Geographic search

North

West East

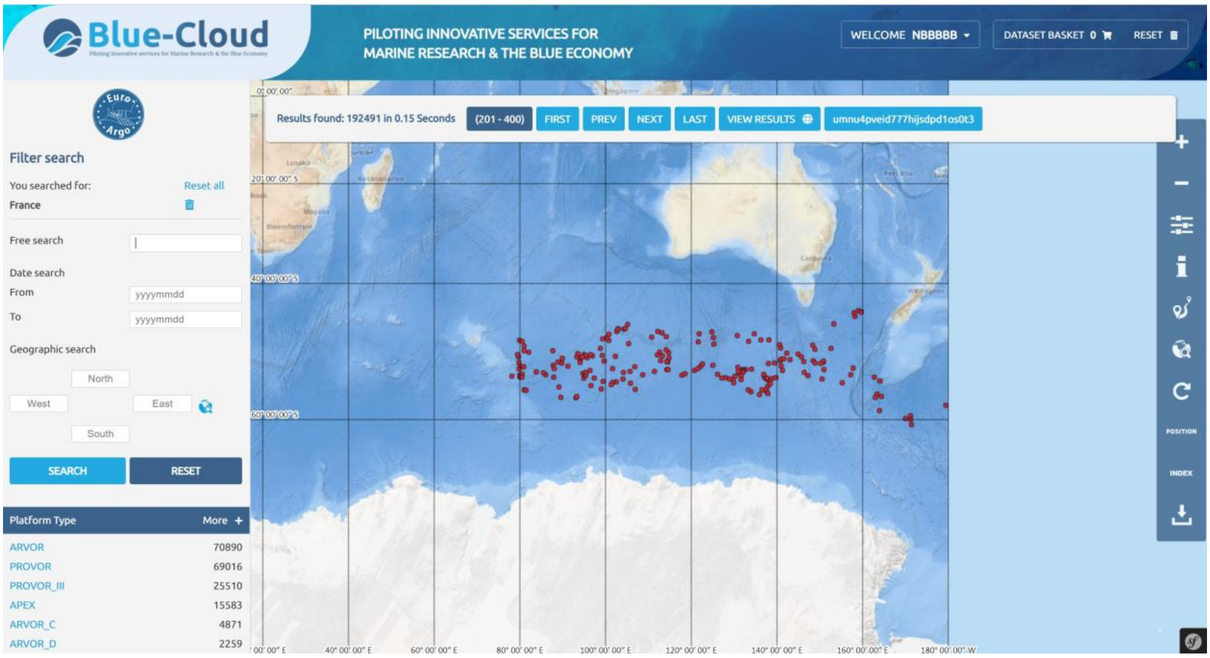
South

Platform Type	More +
APEX	1303991
ARVOR	214163
SOLO	161183
SOLO_II	140264
SOLO_W	135844
PROVOR	117382
NAVIS_A	114053

Results found: 2443475 in 0.07 Seconds

Datacenter	Platform	Platform Code	Transmission System	Cycle	Cycle Start Date
○ CORIOLIS	PROVOR-CST3 DO Profiling Float	6902874	IRIDIUM	43	2019-03-08
○ CORIOLIS	PROVOR-CST3 DO Profiling Float	6902873	IRIDIUM	78	2020-01-26
○ CORIOLIS	ARVOR Profiling Float	6902857	IRIDIUM	63	2020-02-15
○ CORIOLIS	PROVOR-CST3 DO Profiling Float	6902873	IRIDIUM	62	2019-11-07
○ CORIOLIS	PROVOR-CST3 DO Profiling Float	6902868	IRIDIUM	43	2019-08-27
○ CORIOLIS	PROVOR-CST3 DO Profiling Float	6902873	IRIDIUM	46	2019-08-19
○ CORIOLIS	ARVOR Profiling Float	6902857	IRIDIUM	47	2019-09-08
○ CORIOLIS	PROVOR-CST3 DO Profiling Float	6902873	IRIDIUM	30	2019-05-31
○ CORIOLIS	PROVOR-CST3 DO Profiling Float	6902872	IRIDIUM	123	2020-01-27
○ CORIOLIS	PROVOR-CST3 DO Profiling Float	6902872	IRIDIUM	115	2019-12-18
○ CORIOLIS	PROVOR-CST3 DO Profiling Float	6902874	IRIDIUM	13	2018-12-23
○ CORIOLIS	ARVOR Profiling Float	6902857	IRIDIUM	27	2019-02-20

Figure 6.3: Browse of Argo records at level 2 with options for retrieving metadata details of each granule (platform + cycle) and queries at level 2 with extra criteria by means of facets



Results found: 192491 in 0.15 Seconds

Filter search

You searched for: France

Reset all

Free search

Date search

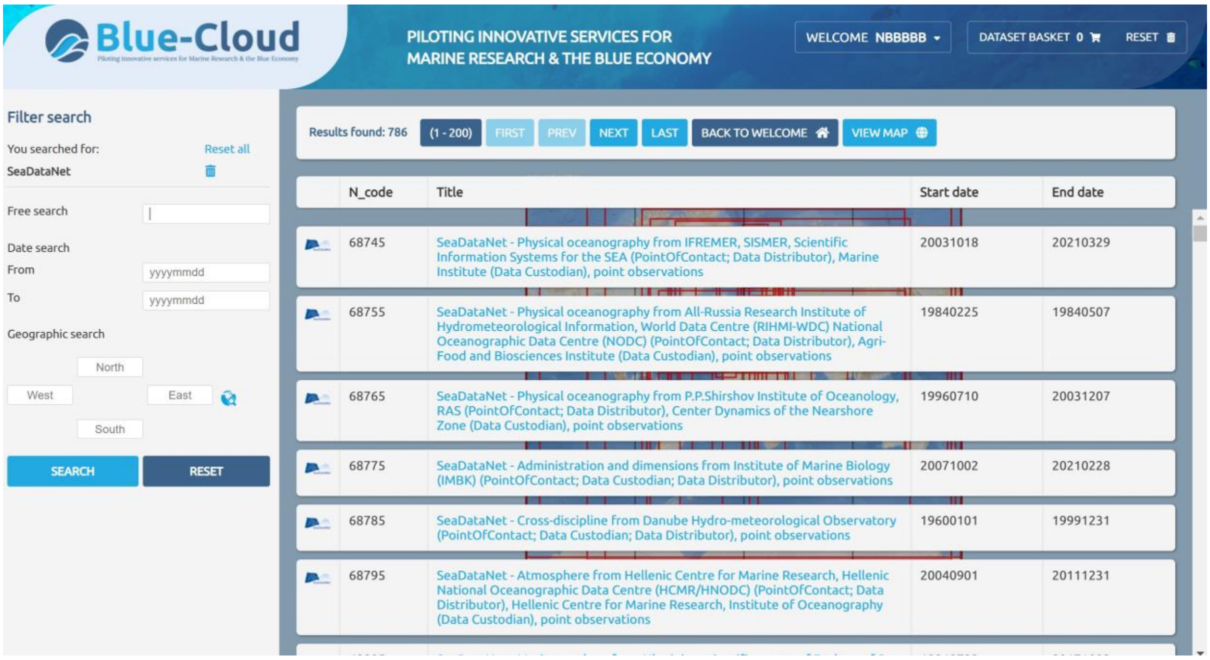
From yyyymmdd

To yyyymmdd

Geographic search

Platform Type	
ARVOR	70890
PROVOR	69016
PROVOR_III	25510
APEX	15583
ARVOR_C	4871
ARVOR_D	2259

Figure 6.4: Browse of Argo records at level 2 on the map with options for retrieving metadata details of each Argo point at granule level on the map and queries at level 2 with extra criteria by means of facets


Blue-Cloud
Piloting innovative services for Marine Research & the Blue Economy
PILOTING INNOVATIVE SERVICES FOR
MARINE RESEARCH & THE BLUE ECONOMY
WELCOME NBBBBB
DATASET BASKET 0
RESET
(1 - 200)
FIRST
PREV
NEXT
LAST
BACK TO WELCOME
VIEW MAP
North
West
East
South
SEARCH
RESET

Filter search

You searched for: SeaDataNet

Reset all

Free search

Date search

From yyyymmdd

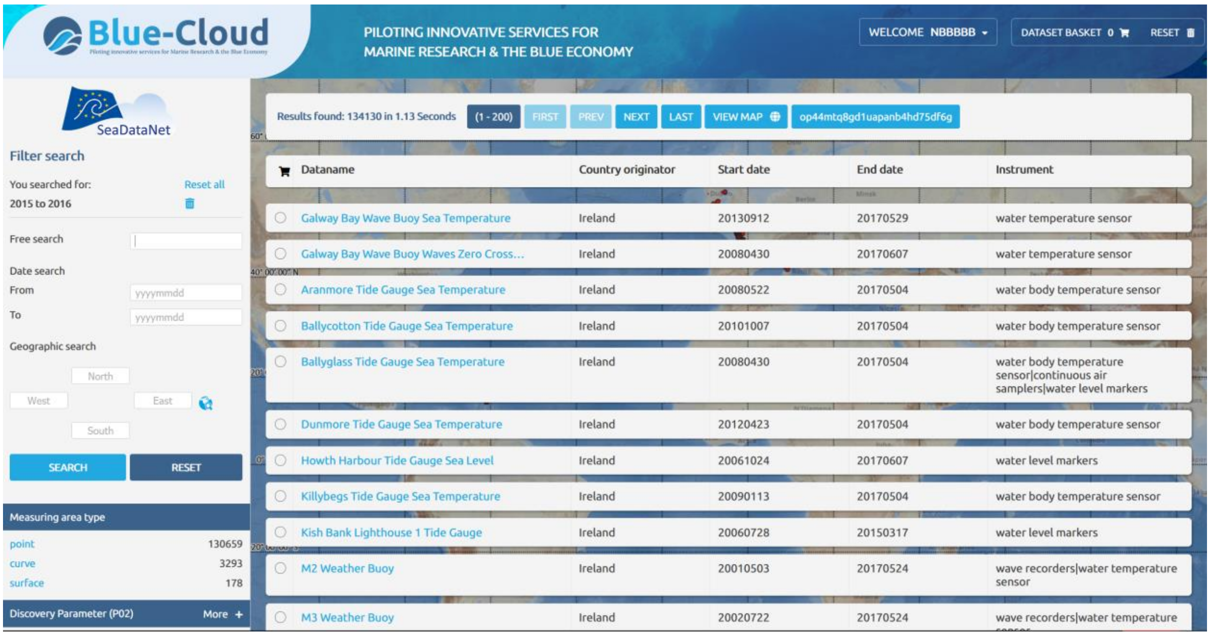
To yyyymmdd

Geographic search

Results found: 786

	N_code	Title	Start date	End date
	68745	SeaDataNet - Physical oceanography from IFREMER, SISMER, Scientific Information Systems for the SEA (PointOfContact; Data Distributor), Marine Institute (Data Custodian), point observations	20031018	20210329
	68755	SeaDataNet - Physical oceanography from All-Russia Research Institute of Hydrometeorological Information, World Data Centre (RIHMI-WDC) National Oceanographic Data Centre (NODC) (PointOfContact; Data Distributor), Agri-Food and Biosciences Institute (Data Custodian), point observations	19840225	19840507
	68765	SeaDataNet - Physical oceanography from P.P.Shirshov Institute of Oceanology, RAS (PointOfContact; Data Distributor), Center Dynamics of the Nearshore Zone (Data Custodian), point observations	19960710	20031207
	68775	SeaDataNet - Administration and dimensions from Institute of Marine Biology (IMBK) (PointOfContact; Data Custodian; Data Distributor), point observations	20071002	20210228
	68785	SeaDataNet - Cross-discipline from Danube Hydro-meteorological Observatory (PointOfContact; Data Custodian; Data Distributor), point observations	19600101	19991231
	68795	SeaDataNet - Atmosphere from Hellenic Centre for Marine Research, Hellenic National Oceanographic Data Centre (HCMR/HNODC) (PointOfContact; Data Distributor), Hellenic Centre for Marine Research, Institute of Oceanography (Data Custodian), point observations	20040901	20111231

Figure 6.5: Browse of SeaDataNet CDI records at level 1 with options for retrieving metadata details of each collection and queries at level 1



Blue-Cloud

PILOTING INNOVATIVE SERVICES FOR MARINE RESEARCH & THE BLUE ECONOMY

WELCOME NBBBBB ▾ | DATASET BASKET 0 | RESET

SeaDataNet

Filter search

You searched for: 2015 to 2016 | Reset all

Free search

Date search
From yyyymmdd
To yyyymmdd

Geographic search
North
West East
South

SEARCH | RESET

Measuring area type

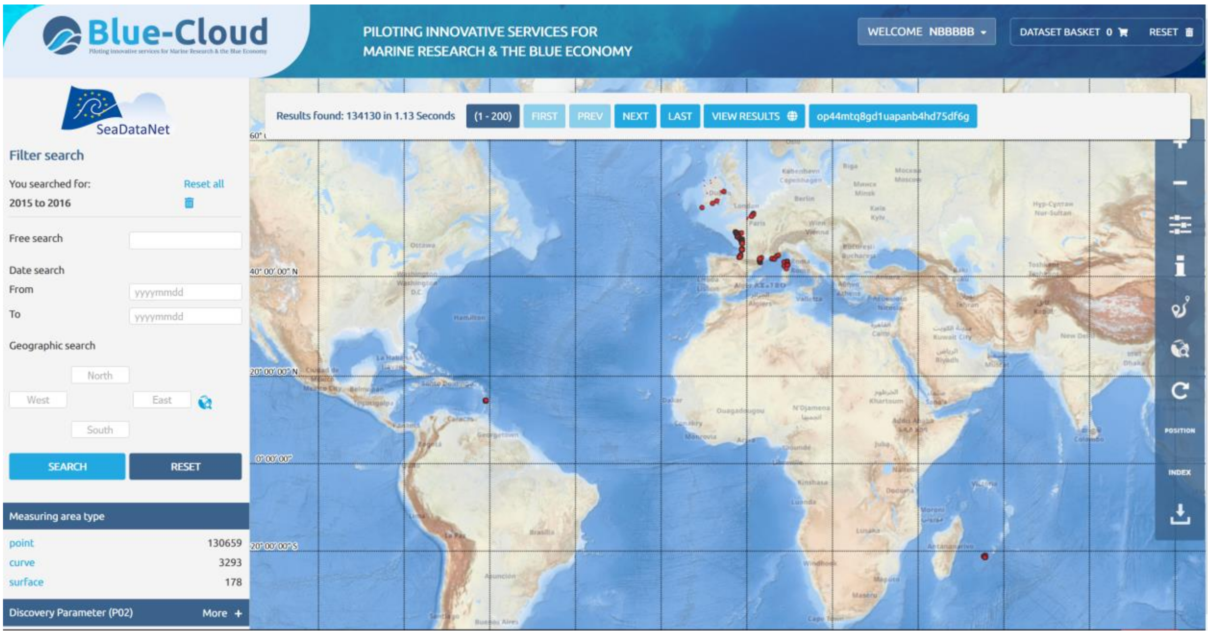
point	130659
curve	3293
surface	178

Discovery Parameter (P02) More +

Results found: 134130 in 1.13 Seconds | (1 - 200) | FIRST | PREV | NEXT | LAST | VIEW MAP | op44mtq8gd1uapanb4hd75df6g

Dataname	Country originator	Start date	End date	Instrument
○ Galway Bay Wave Buoy Sea Temperature	Ireland	20130912	20170529	water temperature sensor
○ Galway Bay Wave Buoy Waves Zero Cross...	Ireland	20080430	20170607	water temperature sensor
○ Aranmore Tide Gauge Sea Temperature	Ireland	20080522	20170504	water body temperature sensor
○ Ballycotton Tide Gauge Sea Temperature	Ireland	20101007	20170504	water body temperature sensor
○ Ballyglass Tide Gauge Sea Temperature	Ireland	20080430	20170504	water body temperature sensor|continuous air samplers|water level markers
○ Dunmore Tide Gauge Sea Temperature	Ireland	20120423	20170504	water body temperature sensor
○ Howth Harbour Tide Gauge Sea Level	Ireland	20061024	20170607	water level markers
○ Killybegs Tide Gauge Sea Temperature	Ireland	20090113	20170504	water body temperature sensor
○ Kish Bank Lighthouse 1 Tide Gauge	Ireland	20060728	20150317	water level markers
○ M2 Weather Buoy	Ireland	20010503	20170524	wave recorders|water temperature sensor
○ M3 Weather Buoy	Ireland	20020722	20170524	wave recorders|water temperature

Figure 6.6: Browse of SeaDataNet CDI records at level 2 with options for retrieving metadata details of each CDI data granule and queries at level 2 with extra criteria by means of facets



Blue-Cloud

PILOTING INNOVATIVE SERVICES FOR MARINE RESEARCH & THE BLUE ECONOMY

WELCOME NBBBBB ▾ | DATASET BASKET 0 | RESET

SeaDataNet

Filter search

You searched for: 2015 to 2016 | Reset all

Free search

Date search
From yyyymmdd
To yyyymmdd

Geographic search
North
West East
South

SEARCH | RESET

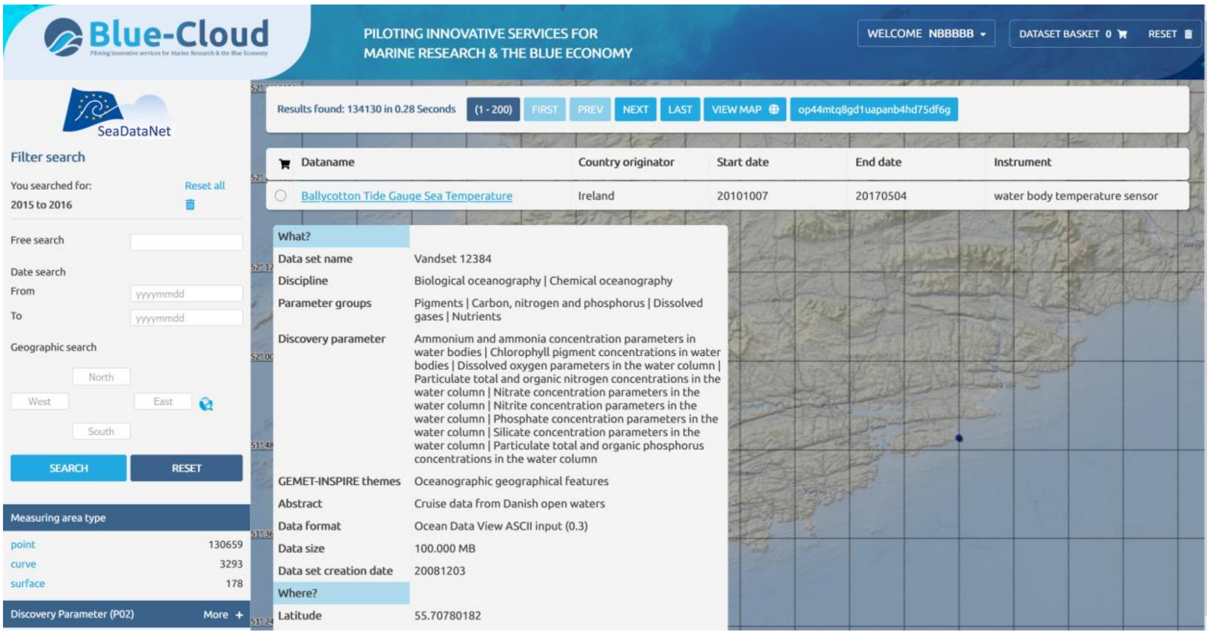
Measuring area type

point	130659
curve	3293
surface	178

Discovery Parameter (P02) More +

Results found: 134130 in 1.13 Seconds | (1 - 200) | FIRST | PREV | NEXT | LAST | VIEW RESULTS | op44mtq8gd1uapanb4hd75df6g

Figure 6.7: Browse of SeaDataNet CDI records at level 2 on the map with options for retrieving metadata details of each CDI record at granule level on the map and queries at level 2 with extra criteria by means of facets


Blue-Cloud
Piloting innovative services for Marine Research & the Blue Economy
PILOTING INNOVATIVE SERVICES FOR
MARINE RESEARCH & THE BLUE ECONOMY
WELCOME NBBBBB
DATASET BASKET 0
RESET
SeaDataNet
(1 - 200)
FIRST
PREV
NEXT
LAST
VIEW MAP
op44mtq8gd1uapanb4hd75df6g
SEARCH
RESET
More +

Filter search

You searched for: 2015 to 2016 Reset all

Free search

Date search
From yyyymmdd
To yyyymmdd

Geographic search
North
West East
South

Measuring area type	
point	130659
curve	3293
surface	178

Discovery Parameter (P02)

Results found: 134130 in 0.28 Seconds

Dataname	Country originator	Start date	End date	Instrument
○ Ballycotton Tide Gauge Sea Temperature	Ireland	20101007	20170504	water body temperature sensor

What?	
Data set name	Vandset 12384
Discipline	Biological oceanography | Chemical oceanography
Parameter groups	Pigments | Carbon, nitrogen and phosphorus | Dissolved gases | Nutrients
Discovery parameter	Ammonium and ammonia concentration parameters in water bodies | Chlorophyll pigment concentrations in water bodies | Dissolved oxygen parameters in the water column | Particulate total and organic nitrogen concentrations in the water column | Nitrate concentration parameters in the water column | Nitrite concentration parameters in the water column | Phosphate concentration parameters in the water column | Silicate concentration parameters in the water column | Particulate total and organic phosphorus concentrations in the water column
GEMET-INSPIRE themes	Oceanographic geographical features
Abstract	Cruise data from Danish open waters
Data format	Ocean Data View ASCII input (0.3)
Data size	100.000 MB
Data set creation date	20081203
Where?	
Latitude	55.70780182

Figure 6.8: Details of a SeaDataNet CDI record at level 2 including location on the map and options for putting in shopping basket and queries at level 2 with extra criteria by means of facets

The background map in the interface is derived from EMODnet Bathymetry and it is the EMODnet Bathymetry World Base Layer (EBWBL), which is made available by OGC WMTS layer. Its GetCapabilities can be found at:
https://tiles.emodnet-bathymetry.eu/