Blue-Cloud

Piloting innovative services for Marine Research & the Blue Economy

D2.2 Blue Data Infrastructures – Services Analysis Report

Work Package	WP2, developing the Blue Cloud discovery and access service and overall Blue Cloud architecture
Lead Partner	MARIS
Lead Author (Org)	MARIS and CNR-IIA
Contributing Author(s)	Dick M.A. Schaap (MARIS), Enrico Boldrini (CNR-IIA), Gilbert Maudire (IFREMER), Thierry Carval (IFREMER), Renaud Dussurget (MOI), Stephane Pesant (EMBL-EBI), Vishnukumar Balavenkataraman Kadhirvelu (EMBL-EBI), Lennert Schepers (VLIZ), Bart VanHoorne (VLIZ), Jean-Olivier Irisson (SU), Benjamin Pfeil (UiB), Steve Jones (UiB)
Reviewers	TCOM members
Due Date	31.10.2020
Submission Date	18.11.2020
Version	1.0

Dissemination Level

X	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	23.10.2020	CNR-IIA	First version
0.2	30.10.2020	MARIS	Second version
0.3	16.11.2020	TCOM members	Internal review
1.0	18.11.2020	MARIS	Final version
1.0	18.11.2020	TRUST-IT	Deliverable submission

Contents

Executive summary

The **Blue Cloud data discovery and access service** will be one of the components of the Blue-Cloud technical framework. It will serve federated discovery and access to blue data infrastructures for external users and also interact with the Blue-Cloud Virtual Research Environment (the component federating computing platforms and analytical services) for populating the VRE data pool. 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. While a further analysis of the architecture and concept for the Blue Cloud data discovery and access service has been described in deliverable D2.6 – Blue Cloud Architecture (1st Release). The Blue Cloud data discovery and access service will provide a common interface, both by web services as by GUI, for discovery and retrieval of data collections from the federated blue data infrastructures. Thereby, conceptually it is planned to set up the query mechanism as a two-step approach, whereby the first step will focus on identifying interesting data collections and products using a common set of search criteria, while the second step will focus on drilling down and sub-setting within the identified collections, in order to get more specific data sets. For the second step, geographic and temporal criteria will be instrumental, next to additional criteria which could be specific per blue data infrastructure. 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. 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.

Implementing this conceptual approach Blue Cloud data discovery and access service will largely depend on the machine-to-machine interfaces of blue data infrastructures, that should be supportive. For that purpose, further analysis activities have been undertaken by MARIS and CNR-IIA for each of the blue data infrastructures concerning the functioning of the existing web services, and in how far these web services are already fit for purpose for the Blue Cloud data discovery and access service. The analyses have been undertaken by try-outs of the earlier given GUI's and web services, and by a series of bilateral web meetings and e-mail communication with the technical contact persons of each of the blue data infrastructures. Moreover, the draft findings and further actions have been presented and discussed at the third TCOM meeting which took place 20 - 21 October 2020. As a follow-up, this draft deliverable D2.2 - Blue Data Infrastructures – Services Analysis Report has been prepared by MARIS and CNR-IIA, describing the results of the analyses and formulating actions for each of the blue data infrastructures, where required. The draft D2.2 has been circulated to the TCOM in order to gather further feedback, and to finalise the draft for submission.

On short term, the formulated actions as listed in Chapter 4 should be followed up by technical staff of relevant blue data infrastructures for making their web services fit for purpose, while the technical team of the Blue-Cloud data discovery and access service should continue their planned developments. The ultimate aim is to establish an operational Blue-Cloud data discovery and access service by M18 (end March 2021). This is only feasible if the web services / APIs are well-functioning and fit-for-purpose at each of the blue data infrastructure.

1 Introduction

The **Blue Cloud data discovery and access service** component will serve federated discovery and access to the following blue data infrastructures:

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

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

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

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 deliverable D2.6 – Blue Cloud Architecture (1st Release), which describes 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. The findings and required follow-up actions will be described in this deliverable D2.2 - Blue Data Infrastructures – Services Analysis Report.

2 Overall concept of Blue Cloud data discovery and access service

The **Blue Cloud Data Discovery and Access service** will facilitate discovery and retrieval of in-situ data, earth observation data, data products, and model output, for external users in stand-alone mode, and for users of the Blue-Cloud 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 overall concept is that the Blue-Cloud Data Discovery and Access service will harvest both metadata and data from the blue data infrastructures by means of web services.

The Blue Cloud data discovery and access service will provide a common interface, both by web services as by GUI, for discovery and retrieval of data collections from the federated blue data infrastructures. The GUI will include facilities for mapping and viewing the locations of data sets, as this will be part of the query dialogue. Moreover, conceptually it is planned to set up the query mechanism as a two-step approach, whereby the first step will focus on identifying interesting data collections and products, using a common set of search criteria, while the second step will focus on drilling down and sub-setting within the identified collections and products in order to get more specific data. For the second step, geographic and temporal criteria will be instrumental, next to additional criteria which could be specific per blue data infrastructure and types of collections. 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 necessary to deal with most of the blue data infrastructures, in particular in cases with observation (raw) data which often can be very large collections with numerous data sets. The second step is then necessary to select a specific geographical area or a specific time slot or a specific variable from that large collection. 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 geographic query options should be supported by a mapping interface, which can be compiled from OGC WMS services to be provided by the blue data infrastructures.

As documented in D2.1, a number of blue data infrastructures are deployed as fully open data repositories with direct download links (data in different formats and standards), while others are configured with a shopping mechanism, featuring user login. Use will be made of existing web services (API's), but where needed, new or adapted API's might need to be defined and agreed together with the technical representatives of the blue data infrastructures. These should then be configured by each to be fit for interacting with the central part of the data brokerage service. The API's should deal with the particulars of the local set-ups and also, they should arrange that data requests can be handled and responded at the agreed aggregation level (collection versus granules).

In the Blue Cloud project use will be made of the GEODAB metadata brokerage service software kit as developed and managed by CNR-IIA. The mappings for the 1st query step are to be made against the common GEODAB metadata model, while the criteria for the second query step will vary between the blue data infrastructures and might also depend on the types of collections. The GEODAB service will be set up by CNR-IIA to generate, maintain, and provide the common Blue-Cloud catalogue in a dynamical way with the latest entries as derived from the blue data infrastructures. Moreover, the GEODAB service will keep track of the 2nd level query profiles, where applicable.

For the data access part of the Blue-Cloud data discovery and access service, a data brokerage service will be developed, integrating the Blue Cloud metadata catalogue, including a shopping mechanism and interfaces, both for human users and machines, to support the actual discovery and retrieval functions. This part will make 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.

The resulting Blue Cloud Data Discovery and Access service should 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 should facilitate managers of blue data infrastructures:

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

The following figure is derived from deliverable D2.6 – Blue Cloud Architecture (first release) and gives an overview of the planned architecture of the Blue Cloud discovery and access service.

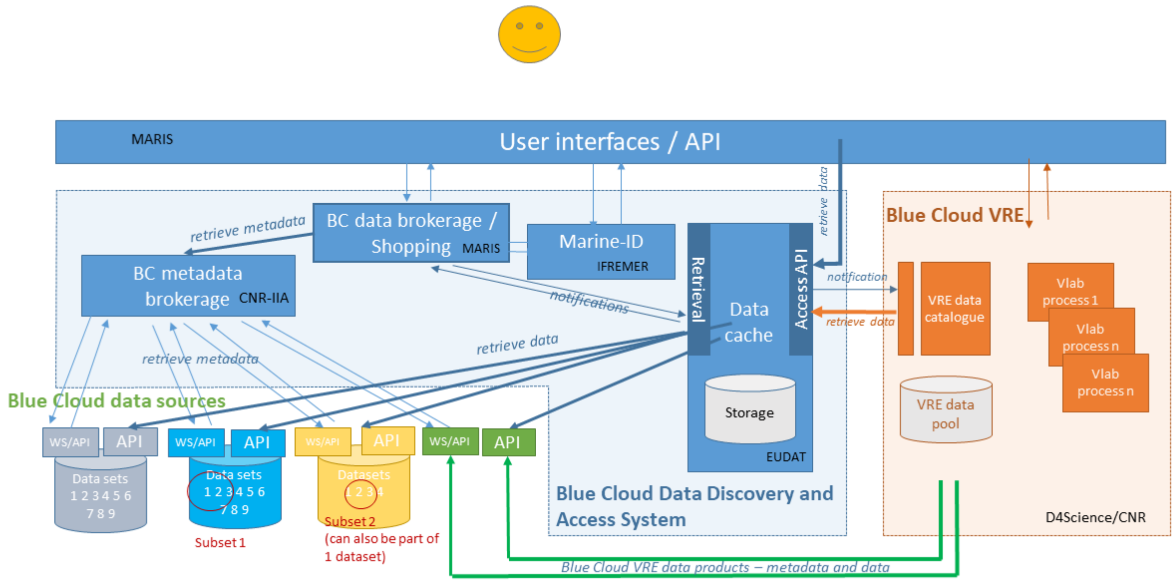


Figure 2.1: Architecture of the Blue Cloud discovery and access service

In summary, 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 will also index and make available selected data products, resulting from the Blue Cloud demonstrator Virtual Labs, to support a wider distribution and publishing.

3 Analyses of web services of Blue Data Infrastructures

Implementing the proposed approach for the Blue Cloud data discovery and access service will largely depend on the machine-to-machine interfaces of the blue data infrastructures, that should be supportive of the conceptual approach. For that purpose, further analysis activities have been undertaken by MARIS and CNR-IIA for each of the blue data infrastructures concerning the functioning of the existing web services, and in how far these web services are already fit for purpose for the Blue Cloud data discovery and access service. The analyses have been undertaken by try-outs of the earlier given GUI's and web services, and by a series of bilateral web meetings and e-mail communication with the technical contact persons of each of the blue data infrastructures. Moreover, the draft findings and further actions have been presented and discussed at the third TCOM meeting which took place 20 - 21 October 2020.

In the following a description is given of how the services from each source have been investigated, in order to find the more efficient way to be discovered by the broker. The typical discovery scenario would involve:

- harvesting metadata at collection level,
- arrange that collection metadata includes info about search criteria for drilling down to granules, and additional links for data access (API), WMS - WFS for granules

And conclusions are formulated about the best way forward and possible actions required from the blue data infrastructures.

3.1 SeaDataNet

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

SeaDataNet publishes datasets and products through two different services:

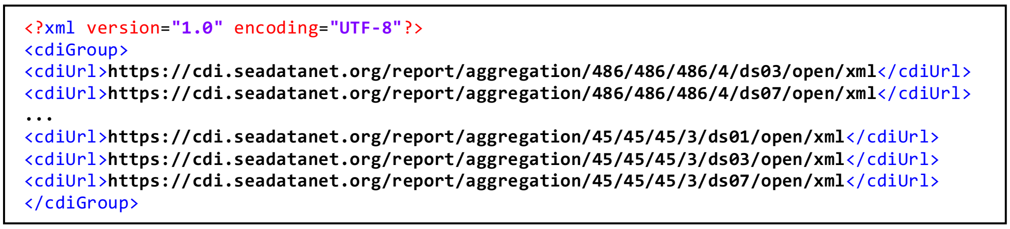
- The **SeaDataNet CDI service** provides harmonized discovery and access to a large volume of marine and ocean data sets (> 2.5 million), both from research and monitoring activities, managed by > 110 data centres;
- The **SeaDataNet products catalogue** publishes SeaDataNet products

3.1.1 The SeaDataNet CDI service

The SeaDataNet CDI service gives access to several million data sets at granule level. To support 1st level queries, a CDI web service has been configured to publish unrestricted SeaDataNet metadata records aggregated as dataset collections. The base inventory is available at the URL (firewall protected): https://cdi.seadatanet.org/report/aggregation/open

The current CDI aggregation works by 3 factors, namely:

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



```
<?xml version="1.0" encoding="UTF-8"?>
<cdiGroup>
<cdiUrl>https://cdi.seadatanet.org/report/aggregation/486/486/486/4/ds03/open/xml</cdiUrl>
<cdiUrl>https://cdi.seadatanet.org/report/aggregation/486/486/486/4/ds07/open/xml</cdiUrl>
...
<cdiUrl>https://cdi.seadatanet.org/report/aggregation/45/45/45/3/ds01/open/xml</cdiUrl>
<cdiUrl>https://cdi.seadatanet.org/report/aggregation/45/45/45/3/ds03/open/xml</cdiUrl>
<cdiUrl>https://cdi.seadatanet.org/report/aggregation/45/45/45/3/ds07/open/xml</cdiUrl>
</cdiGroup>
```

Figure 3.1 CDI collections inventory document

The inventory document (see Figure 3.1) is an XML document with direct links to the individual dataset collections (ca 760 records). Each dataset collection is described with a document compliant with the latest SeaDataNet CDI metadata standard[1].

SeaDataNet CDI is a profile of ISO 19115 metadata, drafted taking into account SeaDataNet community requirements (e.g. in terms of mandatory elements, custom catalogue lists) and compliancy with EU directive INSPIRE[2].

The SeaDataNet CDI collection service is harvested at regular interval of times, and each data collection listed in the inventory can be easily ingested in the broker metadata repository, also based on ISO 19115.

That means that most of the fields can be indexed with little effort (they are to be found in the usual path defined by ISO 19115/ISO 19139). SeaDataNet community elements (into the sdn namespace) have been drafted according to ISO 19115 extension methodology, so it's possible to automatically understand which element they extend, by looking at the *gco:isoType* attribute (e.g. sdn:SDN_DataIdenfication has *gco:isoType="MD_DataIdentification_Type"*). Example given:

Metadata element	**Path**
Title	/gmd:MD_Metadata/ gmd:identificationInfo/sdn:SDN_DataIdentification/gmd:citation/gmd:CI_Citation/gmd:title
Keyword	/gmd:MD_Metadata/gmd:identificationInfo/sdn:SDN_DataIdentification /gmd:descriptiveKeywords/gmd:MD_Keywords/gmd:keyword
Bounding box	/gmd:MD_Metadata/gmd:identificationInfo/sdn:SDN_DataIdentification /gmd:extent/gmd:EX_Extent/gmd:geographicElement/gmd:EX_GeographicBoundingBox

[1] SeaDataNet CDI metadata standard https://www.seadatanet.org/Standards/Metadata-formats/CDI
[2] INSPIRE regulation as regards metadata https://inspire.ec.europa.eu/metadata

Temporal extent	/gmd:MD_Metadata/gmd:identificationInfo/sdn:SDN_DataIdentification/gmd:extent /gmd:EX_Extent/gmd:temporalElement/gmd:EX_TemporalExtent/gmd:extent

The following elements have been indexed to have them available as query filters. Here specific SeaDataNet terms are gathered from correspondent catalogue codelists, also setup according to ISO 19115 extension methodology.

Metadata element	
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
Originator organization	/gmd:MD_Metadata/gmd:distributionInfo/gmd:MD_Distribution/gmd:distributor/gmd:MD_Distributor/gmd:distributorContact/gmd:CI_ResponsibleParty/gmd:organisationName/sdn:SDN_EDMOCode

As a result, the broker publishes the SeaDataNet CDI aggregated dataset through multiple endpoints following the OGC CSW, OGC OpenSearch, and OAI-PMH protocols. The URLs for the CDI service endpoints are as follows:

SeaDataNet CDI OGC CSW endpoint:

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

with SeaDataNet CDI CSW GetCapabilities:

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

SeaDataNet CDI OAI-PMH endpoint:

https://gs-service-production.geodab.eu/gs-service/services/essi/view/seadatanet-open/oaipmh

with SeaDataNet CDI OAI-PMH Identify request:

https://gs-service-production.geodab.eu/gs-service/services/essi/view/seadatanet-open/oaipmh?verb=Identify

OGC OpenSearch description endpoint:

https://gs-service-production.geodab.eu/gs-service/services/essi/view/seadatanet-open/opensearch/description

With OpenSearch based demo portal:

https://gs-service-production.geodab.eu/gs-service/search?view=seadatanet-open

Each metadata record of a CDI collection entry also includes a specific URL for retrieving the associated CDI records at granules level, allowing to drill down, and to derive a specific data access URL:

See e.g.: https://cdi.seadatanet.org/search?step=0082543~0371433~0482543~0116~020ds04

However, currently these functions are only published as GUI and not yet as API's, for which developments are underway by MARIS.

The CDI service has OGC WMS – WFS services are as follows:

WMS service: https://geo-service.maris.nl/seadatanet/wms

WFS service: https://geo-service.maris.nl/seadatanet/wfs

3.1.2 SeaDataNet CDI service – Conclusions and actions

The SeaDataNet CDI service has a web service at collection level which makes it easy to harvest the metadata which are needed for the common metadata profile of the broker service, while also full metadata per collection can be retrieved. The metadata profile of the CDI collection records is the same as that for CDI granule records, which makes it possible to refine and build queries at second level on the CDI granules.

Action MARIS: Further development is required for providing the metadata and data API's for the CDI service at granule level.

Action MARIS: Further development is required for including the specific OGC WMS – WFS service links at collection level, indicating the locations and details of granules.

3.1.3 The SeaDataNet products catalogue

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

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

Currently, 31 records are discoverable from the service, whereby each one is described with an ISO 19115 metadata document.

That means that most of the fields can be indexed with little effort (they are to be found at the usual paths defined by ISO 19115/ISO 19139). Example given:

Metadata element	Path
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

The following are common elements that are useful.

Metadata element	
Parameter	/gmd:MD_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords[gmd:type/gmd:MD_KeywordTypeCode/@codeListValue='parameter']/gmd:keyword/
Instrument	Missing! But not relevant for the products
Platform	Missing! But not relevant for the products
Originator organization	/gmd:MD_Metadata/gmd: identificationInfo /gmd:MD_DataIdentification/gmd:pointOfContact/gmd:CI_ResponsibleParty/gmd:organisationName

As a result, the broker will publish SeaDataNet products catalogue through multiple endpoints following the OGC CSW, OGC OpenSearch, and OAI-PMH protocols. The URLs for the SeaDataNet products catalogue service endpoints are as follows:

SeaDataNet products catalogue OGC CSW endpoint:

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

with SeaDataNet products catalogue CSW GetCapabilities:

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

SeaDataNet products catalogue OAI-PMH endpoint:

https://gs-service-production.geodab.eu/gs-service/services/essi/view/seadatanet-products/oaipmh

with SeaDataNet products catalogue OAI-PMH Identify request:

https://gs-service-production.geodab.eu/gs-service/services/essi/view/seadatanet-products/oaipmh?verb=Identify

OGC OpenSearch description endpoint:

https://gs-service-production.geodab.eu/gs-service/services/essi/view/seadatanet-products/opensearch/description

With OpenSearch based demo portal:

https://gs-service-production.geodab.eu/gs-service/search?view=seadatanet-products

3.1.4 The SeaDataNet products catalogue - Conclusions

The SeaDataNet products catalogue has a web service which makes it easy to harvest the metadata which are needed for the common metadata profile of the broker service, while also full metadata per collection can be retrieved. Each collection metadata record includes a direct link for downloading the product. There is no need for a second query level, as users will be interested in the complete product.

3.2 EMODnet Bathymetry

EMODnet Bathymetry makes use of the SeaDataNet CDI data discovery and access service. Therefore, no separate solutions need to be developed for Blue-Cloud. The highly popular EMODnet Digital Terrain Model (DTM) data product is relevant for Blue-Cloud purposes and can be used through existing OGC web services:

WMS: https://ows.emodnet-bathymetry.eu/wms
WFS: https://ows.emodnet-bathymetry.eu/wfs
WMTS: https://tiles.emodnet-bathymetry.eu
WCS: https://ows.emodnet-bathymetry.eu/wcs

3.3 EMODnet Chemistry

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

3.3.1 EMODnet Chemistry products catalogue

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

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

Currently, 179 records are discoverable from the service, each one is described with an ISO 19115 metadata document.

That means that most of the fields can be indexed with little effort (they are to be found at the usual paths defined by ISO 19115/ISO 19139). Example given:

Metadata element	Path
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

The following are common elements that are useful.

Metadata element	
Parameter	/gmd:MD_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:descriptiveKeywords/gmd:MD_Keywords[gmd:type/gmd:MD_KeywordTypeCode/@codeListValue='parameter']/gmd:keyword/
Instrument	Missing! But not relevant for the products
Platform	Missing! But not relevant for the products
Originator organization	/gmd:MD_Metadata/gmd: identificationInfo /gmd:MD_DataIdentification/gmd:pointOfContact/gmd:CI_ResponsibleParty/gmd:organisationName

As a result, the broker will publish EMODnet Chemistry products through multiple endpoints following the OGC CSW, OGC OpenSearch, and OAI-PMH protocols. The URLs for the EMODnet Chemistry service endpoints are as follows:

EMODnet Chemistry OGC CSW endpoint:

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

with EMODnet Chemistry CSW GetCapabilities:

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

EMODnet Chemistry OAI-PMH endpoint:

https://gs-service-production.geodab.eu/gs-service/services/essi/view/emodnet-chemistry/oaipmh

with EMODnet Chemistry OAI-PMH Identify request:

https://gs-service-production.geodab.eu/gs-service/services/essi/view/emodnet-chemistry/oaipmh?verb=Identify

OGC OpenSearch description endpoint:

https://gs-service-production.geodab.eu/gs-service/services/essi/view/emodnet-chemistry/opensearch/description

With OpenSearch based demo portal:

https://gs-service-production.geodab.eu/gs-service/search?view=emodnet-chemistry

3.3.2 EMODnet Chemistry products catalogue - Conclusions

The EMODnet Chemistry products catalogue has a web service which makes it easy to harvest the metadata which are needed for the common metadata profile of the broker service, while also full metadata per collection can be retrieved. Each collection metadata record includes a direct link for downloading the product. There is no need for a second query level, as users will be interested in the complete product.

3.4 EuroArgo – Argo

EuroArgo operates a number of web services for discovery and access to the ArgoFloat data sets, not only for EuroArgo but for the whole set of ArgoFloats through its GDAC. Very good progress is being made with new advanced services as part of the ENVRI-FAIR, EIOSC-HUB, and EA-RISE projects; these new services are therefore taken into account as candidates for coupling to the Blue-Cloud. Over time there have been circa 20.000 floats of which currently > 3000 active. Each float has a unique Float ID, and each float has collected multiple data sets (= Cycles), each with a unique Cycle ID. Currently, there are more than 2 Million Cycles data sets available in the system. All cycle data sets are made available as NetCDF files. Cycle NetCDFs are replaced by newer versions where needed, but IDs are persistent. Every day circa 300 new cycles are appended to the collection. Monthly extra contributions from other DACs worldwide are added.

Potential available services:

- GDAC floats dashboard https://fleetmonitoring.euro-argo.eu/dashboard
- Data API for machine to machine access https://dataselection.euro-argo.eu/
- GDAC interactive data selection (only for human interaction) http://www.argodatamgt.org/Access-to-data/Argo-data-selection
- GDAC Thredds server http://tds0.ifremer.fr/thredds/catalog/CORIOLIS-ARGO-GDAC-OBS/catalog.html
- OpenSearch service on Argo Data https://opensearch.ifremer.fr/granules.atom?datasetId=argo&startPage=0&timeStart=2020-09-01T00:00:00Z&timeEnd=2020-09-08T23:59:59Z&geoBox=-180.0,-90.0,180.0,90.0
- GDAC ERDDAP data server http://www.ifremer.fr/erddap/tabledap/ArgoFloats.graph
- GDAC rsync service (incremental updates of NetCDF files)

3.4.1 GDAC floats dashboard

This service offers a GUI for discovery, evaluation and download. The GDAC floats dashboard is available at: https://fleetmonitoring.euro-argo.eu/dashboard

The GUI uses a JSON based API that could be leveraged to harvest the interesting metadata.

The API is documented at: https://fleetmonitoring.euro-argo.eu/swagger-ui.html

Record granularity is at the level of floats, having 16487 floats considering also inactive ones.

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

These requests will give back a list of all the available WMO identifiers, unique for each float (= Float ID). Then requests like the following can be made to retrieve full metadata for each float:

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

The JSON metadata result describing each float is quite rich in content, almost comparable to the content of original NetCDF files, containing also detailed information about individual cycles. However, the basic metadata request was instead not sufficient as it misses sensor information, and misses locations of all cycles.

Example mappings for some common metadata fields:

Metadata element	Path
Title	$[*].['platform'].['name'] + $[*].['platform'].['code'] + $[*].['platform'].['description']
Keyword	$[*].['countryCode'] ; $[*].['projectName'] ; $[*].['model'] ; $[*].['maker'] ; $[*].['deployment'].['cruiseName'] ;
Bounding box	Computable from: $[*].['deployment'].['lat'] + $[*].['deployment'].['lon'] (deployment location) + $[*].['locations'].[*].['lat'] + $[*].['locations'].[*].['lon'] (cycle locations)
Temporal extent	$[*].['deployment'].['launchDate'] + $[*].['lastCycle'].['startDate']
Parameter	$[*].['parameters'] (e.g. CTD_PRES, CTD_TEMP, CTD_CNDC, OPTODE_DOXY) - Or, (probably better) - $[*].['variables'] (e.g. SUBSURFACE PRESSURE,SUBSURFACE SALINITY,OXYGEN,SUBSURFACE TEMPERATURE)
Instrument	$[*].['sensors'] (e.g. CTD_PRES, CTD_TEMP, CTD_CNDC, OPTODE_DOXY)
Platform	$[*].['platform'] (e.g. ARVOR-I DO Profiling Float)
Originator organization	$[*].['owner']; $[*].['deployment'].['principalInvestigatorName'] ; $[*].['dataCenter'] ; $[*].[' institution'] This probably would be the best field, but it seems to be often empty. (Data provider to check)

Issues:

- Suggestions are welcome to decide which will be the best mappings for parameter and originator organization; see remarks in table above;
- Desired is adding a link for a WMS showing a map of the trajectory performed by the float. This is already under development by IFREMER, but establishing sufficient performance is challenging considering the large volume of cycle tracks and their updating. As an alternative it could be considered to provide a WMS link per float.

Each cycle is described in the response as well, with the following metadata:

- Cycle number
- Station id
- Station type
- Station direction
- Date

- latitude
- longitude
- grounded
- pmax
- surfacePressure
- surfaceTemperature
- surfaceSalinity
- bottomPressure
- bottomTemperature
- bottomSalinity
- position & date quality

Issues:

- The second level query could be readily executed using one of the above parameters, to retrieve a subset of the collection.
- Download URL for cycles can be obtained using cycle station identifiers. Example given, float 6903238, cycle 30 has station identifiers 60178031 and 60178032.

Correspondent download URL:

http://www.ifremer.fr/co-diffClient/diffusion?restMode=1&ptfCode=6903238&direction=A&format=Netcdf&cvNumber=60178031&lang=en&mode=Argo&formOK=true&formMail=

The response html page contains the FTP address where data will be delivered. E.g.:

ftp://ftp.ifremer.fr/ifremer/coriolis/tmp/co0501/6903238_60178031_20201019142346451.nc

3.4.2 Data API for machine to machine access

Argo Data Selection tool is available at: https://dataselection.euro-argo.eu/

Underneath it is using a convenient API for machine to machine access, documented at: https://dataselection.euro-argo.eu/swagger-ui.html

Another option for second level queries is to make use of this API. Searches for cycles can be made with HTTP-POST requests to:

https://dataselection.euro-argo.eu/api/find-by-search-filtred

In the request body the following parameters can be specified:

- bbox
- country
- cruise
- data center
- deployment year
- group code
- PI
- Parameters

- Platform code
- Platform type
- Telecom
- Temporal extent
- Position & date quality

However, the metadata granules matching in the response will have only the following metadata:

- Id
- cvNumber
- latitude
- longitude
- platform code
- cycleQcState
- level

Issues:

- The response metadata could be improved to contain at least the query parameters. In particular temporal extent would be useful;
- Data download currently requires human intervention (Captcha) and only email from IFREMER is allowed; (try to click NetCDF from the page: https://dataselection.euro-argo.eu/cycle/3044775); this needs to be streamlined

3.4.3 GDAC Thredds server API

The Argo data is originally stored as NetCDF and served also through the GDAC THREDDS server, providing browsing, evaluation and access capabilities through human or machine interaction. Endpoint: http://tds0.ifremer.fr/thredds/catalog/CORIOLIS-ARGO-GDAC-OBS/catalog.html

NetCDF are available according to a folder hierarchy: at the root level there is one folder for **each Argo data center**. Inside each data center folder there is a set of folders, one for **each float** made available by the data center, each one named with its WMO station identifier. Inside each float folder there is a **profile folder**, containing all the profiles acquired by the float and a NetCDF file ending with _meta containing only the metadata about the float. A NetCDF file ending with Rtraj contains the trajectory of the float.

In order to discover Argo resources it is possible to implement harvesting at profile level or float level (probably best option to reduce the number of records). The harvester in this case would crawl the THREDDS hierarchy of folders looking for float metadata. It would be possible to entirely download the interesting NetCDF files in order to read them later with local libraries or use the OPeNDAP service to transmit only the values of interesting variables inside the NetCDF files. The information content will be the most accurate, being NetCDF the original data and metadata format. However, some service errors and slowness during a preliminary test pose some doubts on the possibility of successfully complete the harvesting of the whole archive without errors.

Example mappings for some common metadata fields from NetCDF variables:

Metadata element	Path
Title	PLATFORM_FAMILY + PLATFORM_TYPE + PLATFORM_NUMBER
Keyword	PROJECT_NAME
Bounding box	From associated trajectory NetCDF file (most accurate)
Temporal extent	STARTUP_DATE ; END_MISSION_DATE
Parameter	PARAMETER
Instrument	PARAMETER_SENSOR
Platform	PLATFORM_NUMBER
Originator organization	FLOAT_OWNER ; OPERATING_INSTITUTION ; DATA_CENTRE

Pros: it has a hierarchical structure, easy to harvest and rich in metadata

Cons: metadata for collections is absent.

Suggestions: to improve metadata for folders (e.g. collections). This could be made even automatically by configuring THREDDS to first create aggregated datasets for each folder, and refining the metadata with human intervention later.

3.4.4 Search engine service on Argo Data

The search engine service is published as an OpenSearch endpoint at: https://opensearch.ifremer.fr/granules.atom?datasetId=argo&startPage=0&timeStart=2020-09-01T00:00:00Z&timeEnd=2020-09-08T23:59:59Z&geoBox=-180.0,-90.0,180.0,90.0

The query returns datasets at granule level for a total of 2,263,681 records. The granularity level is at the cycle (many cycles could be available for a single float). The objective of this service is to provide data access to granules. Indeed, each granule points to a NetCDF data file that can be accessed through the following services:

- FTP
- HTTPS
- ARGO Fleet Monitoring
- ARGO Data Selection

Query constraints include:

- Time start
- Time end
- Bounding box
- Pagination

However, the OpenSearch description document endpoint is unavailable. This should be provided to learn more about all the possible query constraints available. Also, only a few metadata elements are present in the OpenSearch response. There are only links to download the granules. From there some metadata could be achieved.

- Extracted directly from NetCDF file (time consuming)
- from ARGO Fleet Monitoring link
- from ARGO Data Selection link

Pros: it's possible both to harvest and to **search real time** for the granules. This latter option is particularly useful, as granules are frequently updated. Download of granules is easily achieved (direct download URL present).

Cons: metadata is not attached to results

3.4.5 GDAC ERDDAP data server API

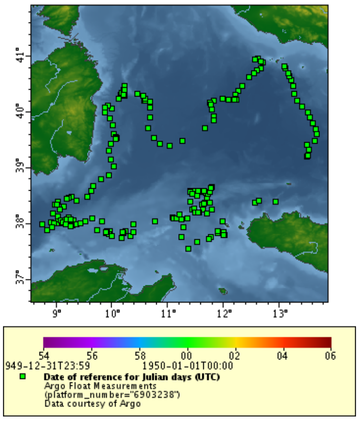
ERDDAP service enables to select subsets of Argo Data through the table DAP form here published: http://www.ifremer.fr/erddap/tabledap/ArgoFloats.html

However, it seems that only a subset of the floats is available through this service as performing a query for distinct platform_number will produce 3090 elements.

Harvesting from this service involves the creation of a query able to produce a metadata table about the floats. In particular, the query should select metadata variables, not data itself. **It seems however that many queries produce service timeouts**. Example given the following query, that should produce a table with distinct values of platforms and the correspondent project names: http://www.ifremer.fr/erddap/tabledap/ArgoFloats.htmlTable?platform_number%2Cproject_name&time%3E=2020-03-16T00%3A00%3A00Z&time%3C=2020-03-23T22%3A58%3A30Z&distinct()

ERDDAP might be used to draw the trajectory of a specific float. This could be used to provide a graphic overview of the collection. E.g. for float 6903238 using the following request:

http://www.ifremer.fr/erddap/tabledap/ArgoFloats.png?longitude,latitude,reference_date_time&platform_number=%226903238%22&.draw=markers&.marker=5%7C5&.color=0x000000&.colorBar=%7C%7C%7C%7C%7C&.bgColor=0xffccccff



3.4.6 EuroArgo – Argo conclusions and actions

After considering all different options, the best option forward would be to use the:

- **GDAC floats dashboard API**

to harvest all the **float metadata (collection level)**, to be used for optimized first level queries.

Second level queries (**granule level**) could be performed using the floats API (retrieving all the cycles for a float, then matching the ones according to user query. Also download URLs would be available. This approach however could be less optimized than applying an alternative approach, using the following service to search for profiles inside a specific float **(granule level)**:

- **GDAC data selection API**

Actions IFREMER:

- Return metadata should be improved (at least with temporal extent of returned cycles);
- Data download should be modified (currently require CAPTCHA and IFREMER e-mail domain). The CAPTCHA is required as it activates processing on the IFREMER HPC server (at DATARMOR) which needs to be secured against robots and hackers. IFREMER will make it possible to authorise subsetting without CAPTCHA when the requests are received from trusted machines, such as Blue-Cloud machines.
- Include a WMS link per float, which should be feasible on short term.

As a final result, the broker will publish the Argo datasets through multiple endpoints following the OGC CSW, OGC OpenSearch, and OAI-PMH protocols. The URLs for the argo service endpoints will be as follows:

Argo OGC CSW endpoint:

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

with Argo CSW GetCapabilities:

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

Argo OAI-PMH endpoint:

https://gs-service-production.geodab.eu/gs-service/services/essi/view/argo/oaipmh

with Argo OAI-PMH Identify request:

https://gs-service-production.geodab.eu/gs-service/services/essi/view/argo/oaipmh?verb=Identify

OGC OpenSearch description endpoint:

https://gs-service-production.geodab.eu/gs-service/services/essi/view/argo/opensearch/description

With OpenSearch based demo portal:

https://gs-service-production.geodab.eu/gs-service/search?view=argo

3.5 EurOBIS – EMODnet Biology

Five different services are reported for discovery of EurOBIS resources from D2.1 – Services Description Report and later additions:

- DCAT service http://ipt.vliz.be/eurobis/dcat
- Geonetwork catalogue at: http://www.emodnet.eu/geonetwork
- Integrated Publishing Toolkit (IPT) at: http://ipt.vliz.be/eurobis/
- IMIS Data Catalog at: https://www.emodnet-biology.eu/data-catalog
- IMIS OAI-PMH (in progress)

After discussion with VLIZ it has been concluded that the best way forward is to make use of the EurOBIS DCAT service as follows.

3.5.1 EurOBIS DCAT service

The EurOBIS DCAT service is published at: http://ipt.vliz.be/eurobis/dcat

It provides a list of all the collections which are hosted at VLIZ and which are public available without restrictions (700+ records). Each data set has a unique 'dasid'. Updating frequency is circa every 3 months, whereby a full overwriting takes place, however persisting the dasid's. For EurOBIS users are interested in only 1 level, the collections, which each can contain millions of observations.

From the dataset links (e.g.

http://ipt.vliz.be/eurobis/resource?r=phytoplankton_in_the_western_north_sea_between_1976_and_1977#Dataset)

it is possible to obtain:

- Metadata description documents, Ecological Metadata Language (EML) encoded (e.g. http://ipt.vliz.be/eurobis/eml?r=phytoplankton_in_the_western_north_sea_between_1976_and_1977#Dataset)
- Download URLs (e.g. http://ipt.vliz.be/eurobis/archive?r=phytoplankton_in_the_western_north_sea_between_1976_and_1977#Dataset)

The direct download URL has been checked to work for all the datasets and is always returning a zipped package.

EML records are being updated approximatively every 3-months. Updates can also be triggered as needed.

EML format needs to be harmonized to ISO 19115 / ISO 19139 through mappings.

Example mappings for some common metadata fields:

Metadata element	Path	Completeness
Title	/eml:eml/dataset/title	100%
Keyword	/eml:eml/dataset/keywordSet/keyword	100%

Bounding box	/eml:eml/dataset/coverage/geographicCoverage/geographicDescription to be converted to BBBOX using Marine regions e.g. http://marineregions.org/mrgid/19828	85%-98%
Temporal extent	/eml:eml/dataset/coverage/temporalCoverage/rangeOfDates	91%
Parameter	Missing! (but present in IMIS)	0%
Instrument	Missing! (but present in IMIS)	0%
Platform	Missing!	0%
Originator organization	/eml:eml/dataset/creator/organizationName	99%

3.5.2 EurOBIS conclusions and actions

After considering all different options, the best option forward is to make use of the **EurOBIS DCAT service** in combination with the **EML records** for full metadata.

Actions VLIZ:

- Bounding box: boundingCoordinates element is present for almost all datasets (98%), but it is often valued with all world extent (-180,+180,-90,+90), while the geographic description (present in 85% of datasets) indicates that the dataset is not global, so the bbox is not correct. Usually, in the geographic description a specific marine region is indicated. Example given, for the dataset at:
  http://ipt.vliz.be/eurobis/eml?r=phytoplankton_in_the_western_north_sea_between_1976_and_1977#Dataset
  Bounding box: -180,+180,-90,+90
  Geographic description: ANE, Belgium, Belgian Continental Shelf (BCS) - http://marineregions.org/mrgid/3293 So, in this case (resolving at https://marineregions.org/mrgid/3293) the correct bbox should be:
  Bounding box: 2.23,3.37,51.08,51.87
  VLIZ should include the correct bounding boxes, using the available information;
- Parameter: the information about parameters is missing in the EML, however the same record from IMIS service has it. Example given:
  EML record:
  http://ipt.vliz.be/eurobis/eml?r=phytoplankton_in_the_western_north_sea_between_1976_and_1977#Dataset
  IMIS record: http://www.vliz.be/imis?dasid=4755&show=json
  The IMIS record has documented the parameters: Density, Phytoplankton abundance
  VLIZ should include the correct parameters, using the available IMIS information;
- Instrument: the information about instruments is missing in the EML, however the same record from IMIS has it. Example given:
  EML record: http://ipt.vliz.be/eurobis/eml?r=arcticssmb#Dataset
  IMIS record: http://www.vliz.be/imis?dasid=533&show=json
  The IMIS record has documented the instrument: Box-corer

VLIZ should include the correct instruments, using the available IMIS information.

As a result, the broker will publish the EurOBIS datasets through multiple endpoints following the OGC CSW, OGC OpenSearch, and OAI-PMH protocols. The URLs for the EurOBIS service endpoints will be as follows:

EurOBIS OGC CSW endpoint:

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

with EurOBIS CSW GetCapabilities:

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

EurOBIS OAI-PMH endpoint:

https://gs-service-production.geodab.eu/gs-service/services/essi/view/eurobis/oaipmh

with EurOBIS OAI-PMH Identify request:

https://gs-service-production.geodab.eu/gs-service/services/essi/view/eurobis/oaipmh?verb=Identify

OGC OpenSearch description endpoint:

https://gs-service-production.geodab.eu/gs-service/services/essi/view/eurobis/opensearch/description

With OpenSearch based demo portal:

https://gs-service-production.geodab.eu/gs-service/search?view=eurobis

3.5.3 EurOBIS GeoNetwork for products

A GeoNetwork catalogue is available at: http://www.emodnet.eu/geonetwork. It is used to publish EurOBIS products and these can be included in Blue-Cloud too. Each record has a description and links to WMS services providing the map products.

Geonetwork publishes two machine-to-machine discovery interfaces:

- an OGC CSW 2.0.2 interface, with ISO Application Profile, at: https://www.emodnet.eu/geonetwork/emodnet/eng/csw
- a custom Geonetwork json based interface at: https://www.emodnet.eu/geonetwork/emodnet/eng/q?

One of these two interfaces can be used to harvest all the metadata records.

Here are sample requests to obtain records from the two interfaces:

- https://www.emodnet.eu/geonetwork/emodnet/eng/csw?service=CSW&request=GetRecords&version=2.0.2&outputFormat=application/xml&outputSchema=http://www.isotc211.org/2005/gmd&ElementSetName=full&resultType=results&typeNames=gmd:MD_Metadata&C

ONSTRAINTLANGUAGE=CQL_TEXT&startPosition=1&maxRecords=10&sourceCatalog=177c58b7-d3e1-4c2b-aac5-7c5a35952b64

- https://www.emodnet.eu/geonetwork/emodnet/eng/q?_content_type=json&facet.q=sourceCatalog%2F177c58b7-d3e1-4c2b-aac5-7c5a35952b64&fast=index&from=1&resultType=details&sortBy=relevance&to=20

The EMODnet Biology records can be retrieved from the Geonetwork OGC CSW via the following filter:

```
<csw:GetRecords xmlns:csw="http://www.opengis.net/cat/csw/2.0.2"
                xmlns:ogc="http://www.opengis.net/ogc"
                xmlns:gmd="http://www.isotc211.org/2005/gmd"
                xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
                xmlns:apiso="http://www.opengis.net/cat/csw/apiso/1.0"
                service="CSW"
                version="2.0.2"
                maxRecords="10"
                startPosition="1"
                resultType="results"
                outputSchema="http://www.isotc211.org/2005/gmd"
                outputFormat="application/xml">
  <csw:Query typeNames="csw:Record">
    <csw:ElementSetName>full</csw:ElementSetName>
       <csw:Constraint version="1.1.0">
            <ogc:Filter xmlns:ogc="http://www.opengis.net/ogc" xmlns:gml="http://www.open
gis.net/gml" xmlns="http://www.opengis.net/ogc">
                <ogc:PropertyIsEqualTo matchCase="false">
                    <ogc:PropertyName>OrganisationName</ogc:PropertyName>
                    <ogc:Literal>Flanders Marine Institute (VLIZ)</ogc:Literal>
                </ogc:PropertyIsEqualTo>
            </ogc:Filter>
        </csw:Constraint>
  </csw:Query>
</csw:GetRecords>
```

The data model mapping will be easy, as the records will be returned already according to ISO 19115 / ISO 19139 (the same data model internally used by the broker). There are however some common fields missing, as noted in the table below.

Example mappings for some common metadata fields:

Metadata element	Path
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	Impossible to identify, as sometimes not present, sometimes the species name is present as a keyword: missing!
Instrument	Missing!
Platform	Missing!
Originator organization	/gmd:MD_Metadata/gmd:identificationInfo/gmd:MD_DataIdentification/gmd:pointOfContact/gmd:CI_ResponsibleParty/gmd:organisationName

3.5.1 EurOBIS products conclusions and actions

The EMODnet Biology products catalogue has a GeoNetwork web service which makes it easy to harvest the metadata which are needed for the common metadata profile of the broker service, while also full metadata per collection can be retrieved. Each collection metadata record includes a direct link for downloading the product. There is no need for a second query level, as users will be interested in the complete product.

Actions VLIZ:

- Updating the publication of its products (also adding new ones), and improving the metadata to the aim of targeting the INSPIRE directive on metadata;

3.6 EcoTaxa

As part of the Blue-Cloud, the ecologically-meaningful data (concentrations of biological organisms per sample) from some datasets in EcoTaxa will be integrated in EurOBIS – EMODnet Biology by a dedicated export API. The coupling to the Blue-Cloud data discovery and access service will then be provided through EurOBIS – EMODnet Biology for this first level query (collections).

EcoTaxa is currently underway with amending their data model to make it easily exportable to the IPT service of EurOBIS and populate it with a number of datasets, such as Tara Oceans. This involves: (i) making their data model compatible at collection-level with the EurOBIS Darwin Core Archive format, (ii) mapping from EcoTaxa taxa names (circa 1700) to WoRMS, which may also require action from VLIZ to include a few more terms in WoRMS.

It should be noted, that not all EcoTaxa records will be publicly available. For that purpose, EcoTaxa is approaching all its data originators (~300 organisations) to get a licence for each data set. The marine collections and associated data with an open licence will be made accessible for the Blue-Cloud.

The second level query approach, with additional criteria, will allow to search within an EcoTaxa collection and to retrieve individual level records and the associated image. It will be supported by a dedicated API that EcoTaxa is currently developing and documenting, based on OpenAPI: https://ecotaxa.obs-vlfr.fr/api/docs

The functions of this API relevant for BlueCloud are the possibility to **browse/search** projects, samples, users, taxa; **query** individual objects matching some criteria; and **export** data**.**

Initial testing could not take place, as the authentication mechanism was not yet functioning in the API. Further communication with EcoTaxa is ongoing to make further testing possible on short term.

3.6.1 EcoTaxa conclusions and actions

EcoTaxa will populate collections (i.e projects) in EurOBIS using the IPT service, which then can be easily harvested by the Blue-Cloud as foreseen for EurOBIS for their full metadata. As second level search, use will be made of the EcoTaxa API to search for granules inside a specific collection. For EcoTaxa, the natural granules are objects and the best aggregation level for those granules will be samples; so ideally, it should facilitate Blue-Cloud users to search using the collection identifier, and a set of additional parameters, e.g.:

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

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

Actions SU:

- Make progress with populating EurOBIS, possibly first with a few relevant and open collections;
- Continue the EcoTaxa API development, considering the Blue-Cloud requirements;
- Support CNR and MARIS during further testing and evaluation;
- Expand the mapping of relevant EcoTaxa metadata fields to EurOBIS ones;
- Arrange choice of licenses with all data providers;

Actions VLIZ:

- Provide support to SU for mapping and population of EurOBIS
- Undertake expanding WoRMS with missing taxa present in EcoTaxa

3.7 ELIXIR-ENA

EMBL-EBI operates APIs for ENA discovery and ENA data retrieval which are suitable endpoints for connecting to the Blue-Cloud data discovery and access service. The ENA system contains many data types / classes and a huge volume of data, which are only partly marine-related. Blue-Cloud should focus on data and information relevant for the marine domain and on data types such as samples and their analyses. A priority list needs to be determined as a next step.

The portal API https://www.ebi.ac.uk/ena/portal/api provides several paths, and the "search" path can be used to create complex discovery queries defined by many parameters such as results format, returned fields, and query constraints. In particular, the returnable fields are described in the "Field definitions" section of the API documentation document available here https://www.ebi.ac.uk/ena/portal/api/doc

The type of query constraints and the type of the returned response fields varies according to the *result type.* The API provides several result types: ***read_study, analysis_study*** are the more appropriate for first level queries (collections) as these provide views of data at the level of "study" which is a collection as defined by a given data provider into the system. Result type ***sample_accession*** will be a good focus for the search at the second level.

3.7.1 Gathering the identifiers of collections

Aim of this step is to gather all the available distinct values of the returnable field *"study_accession".* These values will be used as query constraint in the second step.

The result format shown here for the discovery query of the first step is TSV (Tab Separated Values), although JSON is also available.

For ***analysis_study***, for example, the possible queryable/returnable fields are the following:

```
[
  {
    "columnId": "accession",
    "description": "accession number"
  },
  {
    "columnId": "altitude",
    "description": "Altitude (m)"
  },
  {
    "columnId": "analysis_accession",
    "description": "analysis accession number"
  },
  {
    "columnId": "analysis_alias",
    "description": "submitter's name for the analysis"
  },
  {
    "columnId": "analysis_title",
    "description": "brief sequence analysis description"
  },
  {
    "columnId": "analysis_type",
    "description": "type of sequence analysis"
  },
  {
    "columnId": "assembly_quality",
    "description": "Quality of assembly"
  },
  {
    "columnId": "assembly_software",
    "description": "Assembly software"
  },
  {
    "columnId": "assembly_type",
    "description": "analysis Assembly type"
  },
  {
    "columnId": "binning_software",
    "description": "Binning software"
  },
  {
    "columnId": "bio_material",
    "description": "identifier for biological material including institute
and collection code"
  },
  {
    "columnId": "broker_name",
    "description": "broker name"
  },
  {
    "columnId": "cell_line",
    "description": "cell line from which the sample was obtained"
  },
  {
    "columnId": "cell_type",
    "description": "cell type from which the sample was obtained"
  },
  {
    "columnId": "center_name",
    "description": "Submitting center"
  },
  {
    "columnId": "checklist",
    "description": "checklist name (or ID)"
  },
  {
    "columnId": "collected_by",
    "description": "name of the person who collected the specimen"
  },
  {
    "columnId": "collection_date",
    "description": "date that the specimen was collected"
  },
  {
    "columnId": "completeness_score",
    "description": "Completeness score (%)"
  },
  {
    "columnId": "contamination_score",
    "description": "Contamination score (%)"
  },
  {
    "columnId": "country",
    "description": "locality of sample isolation: country names, oceans or
seas, followed by regions and localities"
  },
  {
    "columnId": "cultivar",
    "description": "cultivar (cultivated variety) of plant from which
sample was obtained"
  },
  {
    "columnId": "culture_collection",
    "description": "identifier for the sample culture including institute
and collection code"
  },
  {
    "columnId": "depth",
    "description": "Depth (m)"
  },
  {
    "columnId": "dev_stage",
    "description": "sample obtained from an organism in a specific
developmental stage"
  },
  {
    "columnId": "ecotype",
    "description": "a population within a given species displaying traits
that reflect adaptation to a local habitat"
  },
  {
    "columnId": "elevation",
    "description": "Elevation (m)"
  },
  {
    "columnId": "environment_biome",
    "description": "Environment (Biome)"
  },
  {
    "columnId": "environment_feature",
    "description": "Environment (Feature)"
  },
  {
    "columnId": "environment_material",
    "description": "Environment (Material)"
  },
  {
    "columnId": "environmental_package",
    "description": "MIGS/MIMS/MIMARKS extension for reporting (from
environment where the sample was obtained)"
  },
  {
    "columnId": "environmental_sample",
    "description": "identifies sequences derived by direct molecular
isolation from an environmental DNA sample"
  },
  {
    "columnId": "experiment_accession",
    "description": "experiment accession number"
  },
  {
      "columnId": "experimental_factor",
      "description": "variable aspects of the experimental design"
    },
    {
      "columnId": "first_public",
      "description": "date when made public"
    },
    {
      "columnId": "germline",
      "description":  "the  sample  is  an  unrearranged  molecule  that  was
inherited from the parental germline"
    },
    {
      "columnId": "host",
      "description": "natural (as opposed to laboratory) host to the organism
from which sample was obtained"
    },
    {
      "columnId": "host_body_site",
      "description": "name of body site from where the sample was obtained"
    },
    {
      "columnId": "host_genotype",
      "description": "genotype of host"
    },
    {
      "columnId": "host_growth_conditions",
      "description": "literature reference giving growth conditions of the
host"
    },
    {
      "columnId": "host_phenotype",
      "description": "phenotype of host"
    },
    {
      "columnId": "host_sex",
      "description": "physical sex of the host"
    },
    {
      "columnId": "host_status",
      "description": "condition of host (eg. diseased or healthy)"
    },
    {
      "columnId": "host_tax_id",
      "description": "NCBI taxon id of the host"
    },
    {
      "columnId": "identified_by",
      "description": "name of the taxonomist who identified the specimen"
    },
    {
      "columnId": "investigation_type",
      "description": "the study type targeted by the sequencing"
    },
    {
      "columnId": "isolate",
      "description": "individual isolate from which sample was obtained"
    },
    {
```

```
    "columnId": "isolation_source",
    "description": "describes the physical, environmental and/or local geographical source of the sample"
  },
  {
    "columnId": "last_updated",
    "description": "date when last updated"
  },
  {
    "columnId": "location",
    "description": "geographic location of isolation of the sample"
  },
  {
    "columnId": "mating_type",
    "description": "mating type of the organism from which the sequence was obtained"
  },
  {
    "columnId": "parent_study",
    "description": "parent study accession number"
  },
  {
    "columnId": "ph",
    "description": "pH"
  },
  {
    "columnId": "pipeline_name",
    "description": "analysis pipeline name"
  },
  {
    "columnId": "pipeline_version",
    "description": "analysis pipeline version"
  },
  {
    "columnId": "project_name",
    "description": "name of the project within which the sequencing was organized"
  },
  {
    "columnId": "protocol_label",
    "description": "the protocol used to produce the sample"
  },
  {
    "columnId": "run_accession",
    "description": "run accession number"
  },
  {
    "columnId": "salinity",
    "description": "Salinity (PSU)"
  },
  {
    "columnId": "sample_accession",
    "description": "sample accession number"
  },
  {
    "columnId": "sample_alias",
    "description": "submitter's name for the sample"
  },
  {
    "columnId": "sample_collection",
    "description": "the method or device employed for collecting the sample"
  },
  {
    "columnId": "sample_material",
    "description": "sample material label"
  },
  {
    "columnId": "sample_title",
    "description": "brief sample title"
  },
  {
    "columnId": "sampling_campaign",
    "description": "the activity within which this sample was collected"
  },
  {
    "columnId": "sampling_platform",
    "description": "the large infrastructure from which this sample was
collected"
  },
  {
    "columnId": "sampling_site",
    "description": "the site/station where this sample was collection"
  },
  {
    "columnId": "secondary_sample_accession",
    "description": "secondary sample accession number"
  },
  {
    "columnId": "secondary_study_accession",
    "description": "secondary study accession number"
  },
  {
    "columnId": "sequencing_method",
    "description": "sequencing method used"
  },
  {
    "columnId": "serotype",
    "description": "serological variety of a species characterized by its
antigenic properties"
  },
  {
    "columnId": "serovar",
    "description": "serological variety of a species (usually a prokaryote)
characterized by its antigenic properties"
  },
  {
    "columnId": "sex",
    "description": "sex of the organism from which the sample was obtained"
  },
  {
    "columnId": "specimen_voucher",
    "description": "identifier for the sample culture including institute
and collection code"
  },
  {
    "columnId": "strain",
    "description": "strain from which sample was obtained"
  },
  {
    "columnId": "study_accession",
    "description": "study accession number"
  },
  {
    "columnId": "study_alias",
    "description": "submitter's name for the study"
  },
  {
    "columnId": "study_title",
    "description": "brief sequencing study description"
  },
  {
    "columnId": "sub_species",
    "description": "name of sub-species of organism from which sample was
obtained"
  },
  {
    "columnId": "sub_strain",
    "description": "name or identifier of a genetically or otherwise
modified strain from which sample was obtained"
  },
  {
    "columnId": "target_gene",
    "description": "targeted gene or locus name for marker gene studies"
  },
  {
    "columnId": "taxonomic_classification",
    "description": "Taxonomic classification"
  },
  {
    "columnId": "taxonomic_identity_marker",
    "description": "Taxonomic identity marker"
  },
  {
    "columnId": "taxonomy",
    "description": "NCBI taxonomic classification"
  },
  {
    "columnId": "temperature",
    "description": "Temperature (C)"
  },
  {
    "columnId": "tissue_lib",
    "description": "tissue library from which sample was obtained"
  },
  {
    "columnId": "tissue_type",
    "description": "tissue type from which the sample was obtained"
  },
  {
    "columnId": "variety",
    "description": "variety (varietas, a formal Linnaean rank) of organism
from which sample was derived"
  }
]
```

In order to retrieve only records related to the marine environment, we must design one or more queries to constraint the whole set of studies (1040 records). A number of options have been investigated, with the following four conditions evaluated as being optimal.

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

Since conditions 1-2 are available directly from services, our focus will initially be on these. Conditions 3-4 are expected to become available during the course of the Blue-Cloud project and we will likely include these later on.

3.7.2 Retrieving metadata of each collection

The second step consists in the discovery of records which provide information about the studies and related samples selected in the first step. Each discovery query selects 10 different studies, using the *"study_accession"* parameter. The selected result format is JSON which is easy to parse if compared with the other available format, TSV.

Example of second step discovery query which selects 2 studies:

https://www.ebi.ac.uk/ena/portal/api/search?fields=sample_description,description,study_title,project_name,lat,lon,location,depth,elevation,altitude,collection_date,first_public,environment_biome,environment_feature,environment_material,environmental_package,investigation_type,country,sample_alias,sequencing_method,scientific_name&format=JSON&limit=10&query=(study_accession="PRJEB11357" OR study_accession="PRJEB12852")&result=analysis_study

This table shows example mappings from the selected returnable fields and common ISO 19115 / ISO 19139 element.

Metadata element	Returnable field	Field description
Title	study_title	Brief sequencing study description
	project_name	Name of the project within which the sequencing was organised
Abstract	description	Brief sequence description

[3] Example query to return read_study records for marine environments:
curl -X POST -H "Content-Type: application/x-www-form-urlencoded" -d 'result=read_study&query=tax_tree(408172)%20OR%20environment_biome%3D%22*marine*%22%20OR%20isolation_source%3D%22*marine*%22&format=json' "https://www.ebi.ac.uk/ena/portal/api/search"

	sample_description	
Keyword	environment_biome	Biomes are defined based on factors such as plant structures, leaf types, plant spacing, and other factors like climate. Examples include: desert, taiga, deciduous woodland, or coral reef
	scientific_name	Scientific name of the organism from which the sample was derived
	environment_feature	Environmental feature level includes geographic environmental features. Examples include: harbor, cliff, or lake
	environment_material	The environmental material level refers to the material that was displaced by the sample, or material in which a sample was embedded, prior to the sampling event. Examples include: air, soil, or water
	environmental_package	MIGS/MIMS/MIMARKS extension for reporting of measurements and observations obtained from one or more of the environments where the sample was obtained
	investigation_type	The study type targeted by the sequencing
	country	Locality of sample isolation: country names, oceans or seas, followed by regions and localities
	sample_alias	Submitter's name for the sample
	sequencing_method	Sequencing method used
Bounding Box	lat	Warning: location is missing from many studies. This being an important discovery field, could this be improved by data provider?
	lon	
Temporal extent	collection_date	Date that the specimen was collected
Vertical extent	depth	The distance below the surface of the water at which a measurement was made or a sample was collected (in metres)
Parameter	scientific_name	
Instrument	Missing! See remarks below	
Platform	Missing! See remarks below	

Originator organization	center_name	

Remarks: Some fields will be missing at the level of analysis_study and read_study as these are aggregates that may combine individual records that cross multiple platforms. Instrument and Platform will be available from lower-level records, such as experiment.

3.7.3 Granule queries

For the second level queries (granules), it is possible to query the same API for the samples associated to a specific study (result type: sample) . The following metadata elements can be used as queryable and returnable:

```
[
  {
    "columnId": "accession",
    "description": "accession number"
  },
  {
    "columnId": "altitude",
    "description": "Altitude (m)"
  },
  {
    "columnId": "assembly_quality",
    "description": "Quality of assembly"
  },
  {
    "columnId": "assembly_software",
    "description": "Assembly software"
  },
  {
    "columnId": "binning_software",
    "description": "Binning software"
  },
  {
    "columnId": "bio_material",
    "description": "identifier for biological material including institute
and collection code"
  },
  {
    "columnId": "broker_name",
    "description": "broker name"
  },
  {
    "columnId": "cell_line",
    "description": "cell line from which the sample was obtained"
  },
  {
    "columnId": "cell_type",
    "description": "cell type from which the sample was obtained"
  },
  {
    "columnId": "center_name",
    "description": "Submitting center"
  },
  {
    "columnId": "checklist",
    "description": "checklist name (or ID)"
  },
  {
    "columnId": "collected_by",
    "description": "name of the person who collected the specimen"
  },
  {
    "columnId": "collection_date",
    "description": "date that the specimen was collected"
  },
  {
    "columnId": "completeness_score",
    "description": "Completeness score (%)"
  },
  {
    "columnId": "contamination_score",
    "description": "Contamination score (%)"
  },
  {
    "columnId": "country",
    "description": "locality of sample isolation: country names, oceans or
seas, followed by regions and localities"
  },
  {
    "columnId": "cultivar",
    "description": "cultivar (cultivated variety) of plant from which
sample was obtained"
  },
  {
    "columnId": "culture_collection",
    "description": "identifier for the sample culture including institute
and collection code"
  },
  {
    "columnId": "depth",
    "description": "Depth (m)"
  },
  {
    "columnId": "description",
    "description": "brief sequence description"
  },
  {
    "columnId": "dev_stage",
    "description": "sample obtained from an organism in a specific
developmental stage"
  },
  {
    "columnId": "ecotype",
    "description": "a population within a given species displaying traits
that reflect adaptation to a local habitat"
  },
  {
    "columnId": "elevation",
    "description": "Elevation (m)"
  },
  {
    "columnId": "environment_biome",
    "description": "Environment (Biome)"
  },
  {
    "columnId": "environment_feature",
    "description": "Environment (Feature)"
  },
  {
    "columnId": "environment_material",
    "description": "Environment (Material)"
  },
  {
    "columnId": "environmental_package",
    "description": "MIGS/MIMS/MIMARKS extension for reporting (from
environment where the sample was obtained)"
  },
  {
    "columnId": "environmental_sample",
    "description": "identifies sequences derived by direct molecular
isolation from an environmental DNA sample"
  },
  {
    "columnId": "experimental_factor",
    "description": "variable aspects of the experimental design"
  },
  {
    "columnId": "first_public",
    "description": "date when made public"
  },
  {
    "columnId": "germline",
    "description": "the sample is an unrearranged molecule that was
inherited from the parental germline"
  },
  {
    "columnId": "host",
    "description": "natural (as opposed to laboratory) host to the organism
from which sample was obtained"
  },
  {
    "columnId": "host_body_site",
    "description": "name of body site from where the sample was obtained"
  },
  {
    "columnId": "host_genotype",
    "description": "genotype of host"
  },
  {
    "columnId": "host_growth_conditions",
    "description": "literature reference giving growth conditions of the
host"
  },
  {
    "columnId": "host_phenotype",
    "description": "phenotype of host"
  },
  {
    "columnId": "host_sex",
    "description": "physical sex of the host"
  },
  {
    "columnId": "host_status",
    "description": "condition of host (eg. diseased or healthy)"
  },
  {
    "columnId": "host_tax_id",
    "description": "NCBI taxon id of the host"
  },
  {
    "columnId": "identified_by",
    "description": "name of the taxonomist who identified the specimen"
  },
  {
    "columnId": "investigation_type",
    "description": "the study type targeted by the sequencing"
  },
  {
    "columnId": "isolate",
    "description": "individual isolate from which sample was obtained"
  },
  {
    "columnId": "isolation_source",
    "description":  "describes  the  physical,  environmental  and/or  local
geographical source of the sample"
  },
  {
    "columnId": "last_updated",
    "description": "date when last updated"
  },
  {
    "columnId": "location",
    "description": "geographic location of isolation of the sample"
  },
  {
    "columnId": "mating_type",
    "description": "mating type of the organism from which the sequence was
obtained"
  },
  {
    "columnId": "ph",
    "description": "pH"
  },
  {
    "columnId": "project_name",
    "description":  "name  of  the  project  within  which  the  sequencing  was
organized"
  },
  {
    "columnId": "protocol_label",
    "description": "the protocol used to produce the sample"
  },
  {
    "columnId": "salinity",
    "description": "Salinity (PSU)"
  },
  {
    "columnId": "sample_accession",
    "description": "sample accession number"
  },
  {
    "columnId": "sample_alias",
    "description": "submitter's name for the sample"
  },
  {
    "columnId": "sample_collection",
    "description": "the method or device employed for collecting the sample"
  },
  {
    "columnId": "sample_material",
    "description": "sample material label"
  },
  {
    "columnId": "sample_title",
    "description": "brief sample title"
  },
  {
    "columnId": "sampling_campaign",
    "description": "the activity within which this sample was collected"
  },
  {
    "columnId": "sampling_platform",
    "description": "the large infrastructure from which this sample was collected"
  },
  {
    "columnId": "sampling_site",
    "description": "the site/station where this sample was collection"
  },
  {
    "columnId": "secondary_sample_accession",
    "description": "secondary sample accession number"
  },
  {
    "columnId": "sequencing_method",
    "description": "sequencing method used"
  },
  {
    "columnId": "serotype",
    "description": "serological variety of a species characterized by its antigenic properties"
  },
  {
    "columnId": "serovar",
    "description": "serological variety of a species (usually a prokaryote) characterized by its antigenic properties"
  },
  {
    "columnId": "sex",
    "description": "sex of the organism from which the sample was obtained"
  },
  {
    "columnId": "specimen_voucher",
    "description": "identifier for the sample culture including institute and collection code"
  },
  {
    "columnId": "strain",
    "description": "strain from which sample was obtained"
  },
  {
    "columnId": "sub_species",
    "description": "name of sub-species of organism from which sample was
obtained"
  },
  {
    "columnId": "sub_strain",
    "description":  "name  or  identifier  of  a  genetically  or  otherwise
modified strain from which sample was obtained"
  },
  {
    "columnId": "target_gene",
    "description": "targeted gene or locus name for marker gene studies"
  },
  {
    "columnId": "taxonomic_classification",
    "description": "Taxonomic classification"
  },
  {
    "columnId": "taxonomic_identity_marker",
    "description": "Taxonomic identity marker"
  },
  {
    "columnId": "taxonomy",
    "description": "NCBI taxonomic classification"
  },
  {
    "columnId": "temperature",
    "description": "Temperature (C)"
  },
  {
    "columnId": "tissue_lib",
    "description": "tissue library from which sample was obtained"
  },
  {
    "columnId": "tissue_type",
    "description": "tissue type from which the sample was obtained"
  },
  {
    "columnId": "variety",
    "description": "variety (varietas, a formal Linnaean rank) of organism
from which sample was derived"
  }
]
```

Spatial (location element) and temporal (collection_date element) attibutes are often missing at granule level, so these cannot be the sole parameters used to refine the second level search. Fields such as taxon and country (which provides in some cases the names of marine areas) should also be considered.

Direct download URL is generally available (download of tar.gz packages from FTP).

3.7.4 ELIXIR-ENA conclusions

Summarizing, collection level records will be returned using ELIXIR-ENA API based upon marine queries that will be designed. The results will be genomics studies of different types. Therefore, we need to proceed with study level using the parent study accessions or individual study accessions. For TARA, PRJEB402 is the parent accession as provided in the example above.

Instrument and platform metadata are not available at collections level, as study records are aggregates and instruments used can be mixed across studies, so values are not propagated to this level. However, instrument metadata are always available for experiment records.

The second level query will be used to retrieve a subset of the interesting collections (the samples acquired at specific locations or for given taxa and experiments). Download of granules will then be possible.

As a result, the broker will publish the ELIXIR-ENA datasets through multiple endpoints following the OGC CSW, OGC OpenSearch, and OAI-PMH protocols. The URLs for the ELIXIR-ENA service endpoints will be as follows:

ELIXIR-ENA OGC CSW endpoint:

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

with ELIXIR-ENA CSW GetCapabilities:

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

ELIXIR-ENA OAI-PMH endpoint:

https://gs-service-production.geodab.eu/gs-service/services/essi/view/elixir-ena/oaipmh

with ELIXIR-ENA OAI-PMH Identify request:

https://gs-service-production.geodab.eu/gs-service/services/essi/view/elixir-ena/oaipmh?verb=Identify

OGC OpenSearch description endpoint:

https://gs-service-production.geodab.eu/gs-service/services/essi/view/elixir-ena/opensearch/description

With OpenSearch based demo portal:

https://gs-service-production.geodab.eu/gs-service/search?view=elixir-ena

Actions EMBL:

- EMBL should check if it can complete the spatial (location element) and temporal (collection_date element) extents that are often missing at granule level, while these should be common parameters useful to refine the second level search. It is noted that the information may not have been provided by the data owners and may not be available, but there are ways around this in some cases.

3.8 EuroBioImaging

As part of EuroBioImaging the BioImage Archive is operated and will be indexed under the EBI Search API (https://www.ebi.ac.uk/ebisearch/overview.ebi/about). This is a recent change and details are not yet available.

Content is only partly marine related. Blue-Cloud should focus on images and databases relevant for the marine domain. This should be analysed as a next step in order to determine if all databases need to be coupled to the Blue-Cloud.

3.8.1 EuroBioImaging conclusions

CNR and MARIS had analysed EuroBioImaging, comprising a number of related components, each with their search and access functionalities. However, EMBL is underway with new developments and these will bring the discovery and access of the BioImage Archive under the EBI Search API. This will take time and details of relevance for the Blue-Cloud cannot be given at present. Therefore, EuroBioImaging will be analysed again in a later stage.

Actions EMBL:

- Keep CNR and MARIS informed about the new developments and how the EBI Search API will work and can be used for the Blue-Cloud interacting with the BioImage Archive. Ultimately, images will be connected to BioSamples records and this path will be possible.
- Include and provide means to restrict queries to marine topics; for this purpose, EMBL is considering a single attribute that will bring together all data sets of aquatic relevance.

3.9 WEkEO

An API can be used to retrieve metadata records by machine interaction. This API is documented here:

https://wekeo-broker.apps.mercator.dpi.wekeo.eu/databroker/ui/

A token should be requested first to use for requests needing authentication, with the following request:

- Destination URL: https://wekeo-broker.apps.mercator.dpi.wekeo.eu/databroker/gettoken
- Method: GET
- Basic authentication (username & password obtained after registering to the portal)

3.9.1 List of available datasets

Different methods can be used to obtain the list of available datasets

1) The following operation can be used to retrieve dataset identifiers:
   - Destination URL: https://wekeo-broker.apps.mercator.dpi.wekeo.eu/databroker/catalogue/datasets
   - Method: GET
   - Header: Authorization (using the token obtained before)

This operation will bring back a list of **319 data identifiers**.

2) Also, the following request can be sent to retrieve all the records in a brief form.
   - Destination URL: https://wekeo-broker.apps.mercator.dpi.wekeo.eu/databroker/datasets?q=&size=1000&page=0
   - Method: GET
   - Header: No Authorization required in this case.

The result is a JSON encoded list of datasets. **935 brief records** can be retrieved. Excerpt:

```
        {
            "abstract": "'''Short description:'''\n\nFor the Global Ocean - The IFREMER CERSAT Global Blended Mean Wind Fields
include wind components (meridional and zonal), wind module, wind stress, and wind/stress curl and divergence. The associated
error estimates are also provided. The estimation of the 6-hourly blended wind products make use of all of the the remotely
sensed surface winds derived from scatterometers and radiometers available at this time (see PUM) and used as observation inputs
for the objective method dealing with the calculation of 6-hourly wind fields over the global oceans. L4 winds are calculated
from L2b products in combination with ERA interim wind analyses from January 1992 onwards. The analysis is performed for each
synoptic time (00h:00; 06h:00; 12h:00; 18h:00 UTC) and with a spatial resolution of 0.25° in longitude and latitude over the
global ocean.",
            "datasetId": "EO:MO:DAT:WIND_GLO_WIND_L4_REP_OBSERVATIONS_012_006:CERSAT-GLO-BLENDED_WIND_L4_REP-V6-
OBS_FULL_TIME_SERIE",
            "previewImage": "https://wekeo-
broker.apps.mercator.dpi.wekeo.eu/previews/EO_MO_DAT_WIND_GLO_WIND_L4_REP_OBSERVATIONS_012_006_CERSAT-GLO-
BLENDED_WIND_L4_REP-V6-OBS_FULL_TIME_SERIE.png",
            "title": "Global Ocean Wind L4 Reprocessed 6 hourly Observations"
        },
```

These are the “datasets” defined at the WEkEO’s Harmonised Data Access (HDA) service:

- In the case of CMEMS, they correspond to “subDatasets”. Note that they have a suffix, such as “:CERSAT-GLO-BLENDED…” in this case.
- In the case of other data providers, there are no “subDatasets”.

It is probably the best option to use this last operation that will bring back more results (also **all the available layers** will be obtained).

3.9.2 Getting metadata of individual datasets

It is possible to retrieve metadata of each record (collection) by id in full details.

1) A public (no authorized) operation exists for retrieving such metadata. The Official HDA endpoint is: https://wekeo-broker.apps.mercator.dpi.wekeo.eu

Example given:

- Destination URL: https://wekeo-broker.apps.mercator.dpi.wekeo.eu/databroker/datasets/EO:ECMWF:DAT:ERA5_HOURLY_VARIABLES_ON_PRESSURE_LEVELS
- Method: GET

The result is a JSON encoded dataset described with full details.

In particular, the following metadata elements will be in general: *abstract, contact, created, datasetId, details, extent, parameters, status, title*

However not all datasets can be retrieved using the public API, some of them are only accessible through the restricted API.

2) An alternative is to use the CSW endpoint: https://wekeo.eu/guide-catalogue-service-for-web-wekeo-csw

With the following GetCapabilities:

https://pn-csw.apps.mercator.dpi.wekeo.eu/elastic-csw/service?service=CSW&request=GetCapabilities

and GetRecords example to retrieve basic metadata of first 100 records:

https://pn-csw.apps.mercator.dpi.wekeo.eu/elastic-csw/service?service=CSW&request=GetRecords&version=2.0.2&ElementSetName=brief&resultType=results&maxRecords=100

Example mappings for some common metadata fields:

Metadata element	Path	Example
Identifier	$[‘datasetid’]	EO:EUM:DAT:SENTINEL-3:SR_1_SRA___
Title	$[‘title’]	SRAL Level 1B - Sentinel-3

Keyword	$[‘parameters’]	"Operational Oceanography", "Climate", "Atmospheric conditions", "Sea Surface Height", "Ecosystems", "Copernicus Online Data Access", "EUMETCast-Satellite", "EUMETCast-Terrestrial", "Ocean", "EUMETCast-Europe", "Level 1 Data", "EUMETSAT Data Centre"
Bounding box	$[‘extent’].[‘bbox’]	POLYGON((-180.0 90.0, 180.0 90.0, 180.0 -90.0, -180.0 -90.0, -180.0 90.0))
Temporal extent	$[‘extent’]].[‘startDate’] (Warning: missing for some datasets!)	2016-12-13 -
Parameter	$[‘details’].[‘instrumentType’] (Warning: missing for some datasets!) Optional for model-based data sets.	ALTIMETRIC
Instrument	$[‘details’].[‘instrument’] (Warning: missing for some datasets!) Optional for model-based data sets	SRAL
Platform	$[‘details’].[‘satellite’] (Warning: missing for some datasets!) Optional for model-based data sets	Sentinel-3
Originator organization	$[‘contact’].[‘ contactUrl’] (Warning: missing for some datasets!)	http://www.eumetsat.int

As a warning, some datasets such as EO:ECMWF:DAT:SEA_ICE_MONTHLY_AND_DAILY_GRIDDED_DATA_1978_PRESENT are missing the highlighted above metadata elements, whereas others have them.

In either ways, the retrieved metadata has granularity level of collections (useful to perform the first step of discovery).

3.9.3 Querying the collections

The querymetadata operation can be used to retrieve collection specific parameters, useful to retrieve the desired granules. Example given:

- Destination URL: https://wekeo-broker.apps.mercator.dpi.wekeo.eu/databroker/querymetadata/EO:ESA:DAT:SENTINEL-1:SAR
- Method: GET
- Header: Authorization (using the token obtained before)

In general, the result is a JSON encoded list of parameters, as in the following excerpt:

```
        "multiStringSelects": null,
        "stringChoices": [
            {
                "comment": "Product Type",
                "details": {
                    "valuesLabels": {
                        "BS": "CARD-BS",
                        "CARD-COH6": "CARD-COH6",
                        "GRD": "GRD",
                        "GRD-COG": "GRD-COG",
                        "OCN": "OCN",
                        "PLANNED": "PLANNED",
                        "RAW": "RAW",
                        "SLC": "SLC"
                    }
                },
                "isRequired": true,
                "label": "Product Type",
                "name": "productType"
            },
            {
                "comment": "Processing Level",
                "details": {
                    "valuesLabels": {
                        "LEVEL1": "LEVEL1",
                        "LEVEL2": "LEVEL2"
                    }
                },
                "isRequired": false,
                "label": "Processing Level",
                "name": "processingLevel"
            },
```

The given collection-specific parameters can be used to formulate a granule query, such as the following.

- Destination URL: https://wekeo-broker.apps.mercator.dpi.wekeo.eu/databroker/datarequest
- Method: GET
- Body: {"datasetId":"EO:ESA:DAT:SENTINEL-1:SAR","boundingBoxValues":[{"name":"bbox","bbox":[0,0,1,1]}],"dateRangeSelectValues":[{

"name":"position","start":"2018-04-30T00:00:00.000Z","end":"2018-05-01T00:00:00.000Z"}],"stringChoiceValues":[{"name":"productType","value":"BS"}]}

- Header: Authorization (using the token obtained before)

The response will provide a job id to be monitored (e.g. jUI49D92DHCPfEZyv38sReBeZPQ), until it is completed.

Attention, sometimes internal server error 500 is returned, other times (probably in case of many or heavy granules) the response time will be unacceptable for web times.

After discussion with MOI, it turned out that during the second level query an actual order is taking place, as this is the expected behavior for "normal" users of the system. However, for Blue-Cloud we will require an extra step between metadata results and the actual data retrieval.

A set of granules (possibly empty) will be ready to be retrieved, using the following request:

- Destination URL: https://wekeo-broker.apps.mercator.dpi.wekeo.eu/databroker/datarequest/jobs/jUI49D92DHCPfEZyv38sReBeZPQ/result?size=10&page=1
- Method: GET
- Header: Authorization (using the token obtained before)

Here is an excerpt of the response, showing one granule:

```
{
            "downloadUri": null,
            "filename": "S1B_IW_GRDH_1SDV_20181223T181558_20181223T181627_014172_01A569_2623.SAFE",
            "order": null,
            "productInfo": {
                "datasetId": "EO:ESA:DAT:SENTINEL-1:SAR",
                "product": "S1B_IW_GRDH_1SDV_20181223T181558_20181223T181627_014172_01A569_2623.SAFE",
                "productEndDate": "2018-12-23T18:16:27.160603Z",
                "productStartDate": "2018-12-23T18:15:58.144981Z"\
            },
            "size": 834961968,
            "url":                                                                                   "4ebe1d79-36b9-53c0-b2c1-
689ebe84de07/S1B_IW_GRDH_1SDV_20181223T181558_20181223T181627_014172_01A569_2623.SAFE"
        },
```

Few metadata elements are present to describe each granule in the response. For the Blue Cloud purpose, it would be useful that at least basic metadata should be valued (bbox, time extent, title, query parameters used in the second step query), in order to guide the user to choose the needed granule(s). However, this is not currently planned by WEkEO.

3.9.4 Ordering the granule

The dataorder operation is available to download the specified granule(s) with the obtained granule url(s).

3.9.5 WEkEO conclusions and actions

As a summary, WEkEO seems to be almost ready for the two-step query mechanism. WEkEO has currently around 200 **datasets** (first-level) and 900+ **subDatasets** (i.e. CMEMS), representing >3000 accessible layers.

For the Blue Cloud purpose, it is required to know how to perform the 2nd level search within a category and get subsets of these data collections as a list of granules (instances) within the selected category. Following MOI, there should be at least 5 profiles related to the 5 main data-access endpoints offered by WEkEO data providers. The table below gives 5 examples of different behavior.

Data Provider	Example
EUMETSAT-CODA	https://wekeo-broker.apps.mercator.dpi.wekeo.eu/databroker/querymetadata/EO:EUM:DAT:SENTINEL-3:SR_1_SRA___
ECMWF-C3S	https://wekeo-broker.apps.mercator.dpi.wekeo.eu/databroker/querymetadata/EO:ECMWF:DAT:ERA5_HOURLY_VARIABLES_ON_PRESSURE_LEVELS
ECMWF-CAMS	https://wekeo-broker.apps.mercator.dpi.wekeo.eu/databroker/querymetadata/EO:ECMWF:DAT:CAMS_GLOBAL_REANALYSIS_EAC4
CMEMS-Motu	https://wekeo-broker.apps.mercator.dpi.wekeo.eu/databroker/querymetadata/EO:MO:DAT:MEDSEA_REANALYSIS_BIO_006_008:sv03-med-ogs-car-rean-(m
CMEMS-FTP	https://wekeo-broker.apps.mercator.dpi.wekeo.eu/databroker/querymetadata/EO:MO:DAT:GLOBAL_ANALYSIS_FORECAST_PHY_001_024:global-analysis-forecast-phy-001-024-hourly-t-u-v-ssh

Basically, mainly *multiStringSelects* & *stringChoices* will differ in shape and will be more or less related to individual layers and additional data selection criterias, while *boundingBoxes*, *dateRangeSelects* and *datasetId* will be more consistent depending on request.

Concluding, there are still a number of additional requirements from Blue-Cloud to make it fit for the Blue-Cloud 2 step query approach, followed by downloading.

Once these requirements can be fulfilled, then the broker will publish WEkEO datasets through multiple endpoints following the OGC CSW, OGC OpenSearch, and OAI-PMH protocols. The URLs for the WEkEO service endpoints will be as follows:

WEkEO OGC CSW endpoint:

https://gs-service-production.geodab.eu/gs-service/services/essi/view/wekeo/csw

with WEkEO CSW GetCapabilities:

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

WEkEO OAI-PMH endpoint:

https://gs-service-production.geodab.eu/gs-service/services/essi/view/wekeo/oaipmh

with WEkEO OAI-PMH Identify request:

https://gs-service-production.geodab.eu/gs-service/services/essi/view/wekeo/oaipmh?verb=Identify

OGC OpenSearch description endpoint:

https://gs-service-production.geodab.eu/gs-service/services/essi/view/wekeo/opensearch/description

With OpenSearch based demo portal:

https://gs-service-production.geodab.eu/gs-service/search?view=wekeo

Actions MOI:

- Analyse whether it is possible to extend the WEkEO metadata - data HDA API with the following functions, and if so, formulate a short-term development plan:
  1) to harvest an index of the data collections (circa 900) with their full metadata (collection level) and indication to which category (search criteria profile) they belong;
  2) to harvest the resulting full metadata of the identified data granules to show to Blue-Cloud users before ordering data;
  3) to provide instructions how the current HDA API allows to download granules using the dataorder endpoint, with each query corresponding to a granule data set as found in the results list (see 2)
- Provide more information about the master metadata table or relational model of tables that the WEkEO system is using in the background to power the 2 level queries in WEkEO and its data retrieval.

3.10 ICOS – Marine

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

3.10.1 ICOS Data portal

ICOS Data portal (https://data.icos-cp.eu/portal/)

is making use underneath of a SPARQL endpoint based API, that can be leveraged to harvest the desired collection metadata.

The filters that should be set for marine observation data are the following:

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

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

```
# listFilteredDataObjects

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

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

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

where {

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

         ?dobj cpmeta:hasObjectSpec ?spec .

         ?dobj cpmeta:hasSizeInBytes ?size .

         ?dobj cpmeta:hasName ?fileName .

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

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

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

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

}

order by desc(?submTime)

offset 0 limit 61
```

About 100+ records will be retrieved by the query.

Then, it's possible to select each item, with requests such as: https://meta.icos-cp.eu/objects/OYFgMcIfy0zoH4M4EV-T6bQF?format=json

The returned JSON document describes each item with increased details.

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

Metadata element	Path	Example
Identifier	$[‘PID’]	11676/OYFgMcIfy0zoH4M4EV-T6bQF
Title	$[‘citationString’]	van Heuven, S. and Hoppema, M.: Underway physical oceanography and carbon dioxide measurements during POLARSTERN cruise PS92, , doi:10.1594/PANGAEA.865497, 2016.
Keyword	$[‘specification’].[‘keywords’]	SOCAT
Bounding box	$[‘coverageGeoJson’]	[[7.5387, 54.1229], [6.7219, 55.1694], [5.7457, 57.0237], [3.5466, 58.8353], [3.2257, 60.7957], [4.0068, 69.9087], [5.156, 69.9888], [13.9602, 70.2504], [11.9538, 73.8058], [8.8047, 77.9446], [8.0892, 79.9602], [19.9116, 81.0079], [17.5095, 81.4375], [21.6751, 81.0752], [13.603, 81.8157], [7.2697, 82.2138], [13.6384, 81.3086], [9.1031, 80.5491], [10.7622, 80.0297], [9.2077, 79.1379]]
Temporal extent	$[‘specificInfo’].[‘acquisition’].[‘interval’].[‘start’] & $[‘specificInfo].[‘acquisition’].[‘interval’].[‘stop’]	2015-05-19T17:22:05Z 2015-06-27T10:53:54Z
Parameter	$[‘specificInfo].[‘columns’].[‘label’]	Depth water [m]
Instrument	Missing!	
Platform	$[‘specificInfo].[‘acquisition’].[‘station’].[‘name’]	RV Polarstern
Originator organization	$[‘specificInfo].[‘productionInfo].[‘creator’].[‘creator’].[‘name’]	Surface Ocean CO2 Atlas (SOCAT)

It is also possible to download the data after having discovered it, with the exception of raw data.

The protocol for download uses HTTP POST requests and SPARL queries to add items to a cart and then proceed to download. User registration is also supported to keep track of orders. The Handle PID gives the Data ID of each set.

3.10.2 SOCAT portal

The SOCAT portal offers tools for interactively view the SOCAT data products and the download of synthesis and gridded files. An ERRDAP service can be leveraged to harvest the metadata content at: https://ferret.pmel.noaa.gov/socat/erddap/tabledap/socat_v2020_fulldata.subset

The SOCAT portal has multiple data product versions. It is best to focus on the latest version, which currently is the 2020 version.

About 6400+ records are then available from the ERDDAP table dap- example mappings for some common metadata fields from the tabledap response:

Metadata element	Path	Example
Identifier	Expocode	06AQ19860627
Title	dataset_name	ANT_V_Leg_2
Keyword		
Bounding box	geospatial_lon_min & geospatial_lon_max & geospatial_lat_min & geospatial_lat_max &	-60 ; 18 ; -69 ; -34
Temporal extent	time_coverage_start & time_coverage_end	1986-06-27T00:00:00Z ; 1986-09-17T23:59:59Z
Parameter	Missing, but see note below	
Instrument	Missing!	
Platform	platform_name	
Originator organization	Organization	

Parameter information for each dataset is missing from the tabledap subset. Two options are possible:

1) Add the information in the tabledap response, which is the preferred option;
2) Obtain the missing information from the data access table “SOCAT v2020 Data Collection”, available here:
https://ferret.pmel.noaa.gov/socat/erddap/tabledap/socat_v2020_fulldata.html
There are columns for each of the variables presented in SOCAT. For each dataset (identified by a unique expocode) several queries could be issued (one for each variable), to find out if for the specific dataset, the given variable is present. This seems a sort of workaround that extracts metadata while actually accessing the data, as it will be quite time consuming, but feasible.
Actually, the interesting variables would be only the 5 ones here in bold, so only 5 requests for each granule would be needed to retrieve the missing parameter metadata:
   - **sal (salinity, PSU)**
   - Temperature_equi (equilibrator chamber temperature, degrees C)
   - **temp (sea surface temperature, degrees C)**
   - Temperature_atm (sea-level air temperature, degrees C)

- Pressure_equi (equilibrator chamber pressure, hPa)
- **Pressure_atm (sea-level air pressure, hPa)**
- xCO2_water_equi_temp_dry_ppm (umol/mol)
- xCO2_water_sst_dry_ppm (umol/mol)
- xCO2_water_equi_temp_wet_ppm (umol/mol)
- xCO2_water_sst_wet_ppm (umol/mol)
- pCO2_water_equi_temp (uatm)
- pCO2_water_sst_100humidity_uatm (uatm)
- fCO2_water_equi_uatm (uatm)
- fCO2_water_sst_100humidity_uatm (uatm)
- xCO2_atm_dry_actual (actual air xCO2 dry, umol/mol)
- xCO2_atm_dry_interp (interpolated air xCO2 dry, umol/mol)
- pCO2_atm_wet_actual (actual air pCO2 wet, uatm)
- pCO2_atm_wet_interp (interpolated air pCO2 wet, uatm)
- fCO2_atm_wet_actual (actual air fCO2 wet, uatm)
- fCO2_atm_wet_interp (interpolated air fCO2 wet, uatm)
- delta_xCO2 (water xCO2 minus atmospheric xCO2, umol/mol)
- delta_pCO2 (water pCO2 minus atmospheric pCO2, uatm)
- delta_fCO2 (water fCO2 minus atmospheric fCO2, uatm)
- xH2O_equi
- relative_humidity
- specific_humidity
- ship_speed (measured ship speed, knots)
- ship_dir (ship direction, degrees)
- wind_speed_true (true wind speed, m/s)
- wind_speed_rel (relative wind speed, m/s)
- wind_dir_true (true wind direction, degrees)
- wind_dir_rel (relative wind direction, degrees)
- **WOCE_CO2_water (WOCE flag for aqueous CO2)**
- WOCE_CO2_atm (WOCE flag for atmospheric CO2)
- woa_sss (salinity from World Ocean Atlas, PSU)
- pressure_ncep_slp (sea level air pressure from NCEP/NCAR reanalysis, hPa)
- fCO2_insitu_from_xCO2_water_equi_temp_dry_ppm (uatm)
- fCO2_insitu_from_xCO2_water_sst_dry_ppm (uatm)
- fCO2_from_pCO2_water_water_equi_temp (uatm)
- fCO2_from_pCO2_water_sst_100humidity_uatm (uatm)
- fCO2_insitu_from_fCO2_water_equi_uatm (uatm)
- fCO2_insitu_from_fCO2_water_sst_100humidty_uatm (uatm)
- fCO2_from_pCO2_water_water_equi_temp_ncep (uatm)
- fCO2_from_pCO2_water_sst_100humidity_uatm_ncep (uatm)
- fCO2_insitu_from_xCO2_water_equi_temp_dry_ppm_woa (uatm)
- fCO2_insitu_from_xCO2_water_sst_dry_ppm_woa (uatm)
- fCO2_insitu_from_xCO2_water_equi_temp_dry_ppm_ncep (uatm)
- fCO2_insitu_from_xCO2_water_sst_dry_ppm_ncep (uatm)

- fCO2_insitu_from_xCO2_water_equi_temp_dry_ppm_ncep_woa (uatm)
- fCO2_insitu_from_xCO2_water_sst_dry_ppm_ncep_woa (uatm)
- **fCO2_recommended (uatm)**
- delta_temp (Equilibrator Temp - SST, degrees C)
- calc_speed (calculated ship speed, knots)
- etopo2 (bathymetry from ETOPO2, meters)
- gvCO2 (GlobalView xCO2, umol/mol)

The data access table can be used to download individual SOCAT granules, by subsetting as desired along one or more dimensions (column names).

3.10.3 ICOS-Marine conclusions and actions

Summarizing, the following services could be used for discovery:

• ICOS Data portal, for a total of 100+ collections from ICOS. Raw data are not downloadable; the focus should be on L1 and L2 level data (about 1049 datasets).

• SOCAT portal, for its latest product consisting of a harmonised and quality-controlled data collection of circa 6000+ granules

For querying SOCAT, both the ERDDAP API and the ICOS API should be used. SOCAT has more delayed mode data and ICOS more recent. SOCAT is the next level in the data life cycle by adding an external QC.

When using ERDDAP, the most relevant query fields are:

- expocode (unique identifier of a dataset)
- socat_doi (DOI of the SOCAT-enhanced dataset)
- qc_flag (QC assessment of the dataset)
- year (sample year)
- month (sample month of year)
- day (sample day of month)
- hour (sample hour of day)
- minute (sample minute of hour)
- second (sample second of minute)
- longitude (degrees_east)
- latitude (degrees_north)
- depth (sample depth, m)
- sal (salinity, PSU)
- temp (sea surface temperature, degrees C)
- Pressure_atm (sea-level air pressure, hPa)
- WOCE_CO2_water (WOCE flag for aqueous CO2)
- fCO2_recommended (uatm)

As a result, the broker will publish the ICOS-Marine datasets through multiple endpoints following the OGC CSW, OGC OpenSearch, and OAI-PMH protocols. The URLs for the ICOS-Marine service endpoints will be as follows:

ICOS-Marine OGC CSW endpoint:

https://gs-service-production.geodab.eu/gs-service/services/essi/view/icos-marine/csw

with ICOS-Marine CSW GetCapabilities:

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

ICOS-Marine OAI-PMH endpoint:

https://gs-service-production.geodab.eu/gs-service/services/essi/view/icos-marine/oaipmh

with ICOS-Marine OAI-PMH Identify request:

https://gs-service-production.geodab.eu/gs-service/services/essi/view/icos-marine/oaipmh?verb=Identify

OGC OpenSearch description endpoint:

https://gs-service-production.geodab.eu/gs-service/services/essi/view/icos-marine/opensearch/description

With OpenSearch based demo portal:

https://gs-service-production.geodab.eu/gs-service/search?view=icos-marine

Actions UiB:

- Complete the missing metadata as highlighted.

4 Conclusions, actions and planned follow-up

The pilot Blue-Cloud project aims at federating initially in total 10 blue data infrastructures. Each of these existing infrastructures have been described in this deliverable D2.1, in particular with a focus on their current data discovery and access mechanisms, if existing. This initial description and evaluation have been followed by a deeper analysis, detailing the technical specifications of the Blue-Cloud data discovery and access service in deliverable D2.6, and the fitness for purpose of the web services of each of the blue data infrastructures and possible developments required at and by each of the blue data infrastructures in this deliverable D2.2.

D2.2 should be followed by further development and implementation activities at central level and at each of the blue data infrastructures. The technicians of the blue data infrastructures should give a follow-up to the actions as formulated on short term (November – December 2020) as the further development of the Blue Cloud Data Discovery and Access service will depend on having well-functioning web service / APIs at each of the blue data infrastructures, that are fit for the agreed search and access concept. This is a prerequisite for allowing to make further progress.

Actions for each blue data infrastructure are summarized in the following table.

Blue data infrastructure	Operator	Actions
SeaDataNet CDI service	**MARIS**	• Further development is required for providing the metadata and data API's for the CDI service at granule level. • Further development is required for including the specific OGC WMS – WFS service links at collection level, indicating the locations and details of granules.
EuroArgo – Argo	**IFREMER**	• Return metadata should be improved (at least with temporal extent of returned cycles); • Data download should be modified (currently require CAPTCHA and IFREMER e-mail domain). The CAPTCHA is required as it activates processing on the IFREMER HPC server (at DATARMOR) which needs to be secured against robots and hackers. IFREMER will make it possible to authorise subsetting without CAPTCHA when the requests are received from trusted machines, such as Blue-Cloud machines. • Include a WMS link per float, which should be feasible on short term.
EurOBIS – EMODnet Biology	**VLIZ**	• Updating the publication of its products (also adding new ones), and improving the metadata to the aim of targeting the INSPIRE directive on metadata;

EcoTaxa	**SU**	• Make progress with populating EurOBIS, possibly first with a few relevant and open collections; • Continue the EcoTaxa API development, considering the Blue-Cloud requirements • Support CNR and MARIS during further testing and evaluation; • Expand the mapping of relevant EcoTaxa metadata fields to EurOBIS ones; • Arrange choice of licenses with all data providers
	VLIZ	• Provide support to SU for mapping and population of EurOBIS • Undertake expanding WoRMS with missing taxa present in EcoTaxa
ENA	**EMBL**	• check if EMBL can complete the spatial (location element) and temporal (collection_date element) extents that are often missing at granule level, while these should be common parameters useful to refine the second level search. It is noted that the information may not have been provided by the data owners and may not be available, but there are ways around this in some cases.
EuroBioImaging	**EMBL**	• Keep CNR and MARIS informed about the new developments and how the EBI Search API will work and can be used for the Blue-Cloud interacting with the BioImage Archive. Ultimately, images will be connected to BioSamples records and this path will be possible. • Include and provide means to restrict queries to marine topics; for this purpose, EMBL is considering a single attribute that will bring together all data sets of aquatic relevance.
WEkEO	**MOI**	• Analyse further whether it is possible to extend the WEkEO metadata - data HDA API with the following functions, and if so, formulate a short-term development plan: • to harvest an index of the data collections (circa 900) with their full metadata (collection level) and indication to which category (search criteria profile) they belong; • to harvest the resulting full metadata of the identified data granules to show to Blue-Cloud users before ordering data; • to provide instructions how the current HDA API allows to download granules using the dataorder endpoint, with each query corresponding to a granule data set as found in the results list • Provide more information about the master metadata table or relational model of tables that the WEkEO system

		is using in the background to power the 2 level queries in WEkEO and its data retrieval.
ICOS-Marine	**UiB**	- Complete the missing metadata as highlighted.

Table: Summary of actions per blue data infrastructure for web service / APIs

At a central level, CNR-IIA will configure and deploy their GEODAB service to function as Blue-Cloud metadata brokerage service, installing adaptors for each of the Data & Access services following the earlier mappings, and then initialising dynamic harvesting of the metadata for data collections. The resulting Blue-Cloud metadata catalogue will be dynamic and operationally published by CNR-IIA as CSW, OAI-PMH and SPARQL services, on top of which MARIS will develop and deploy the Blue-Cloud catalogue GUI. This should support both the first level searches at collections level and the second level searches at granule level.

As a next step, MARIS will develop an operational Blue-Cloud data brokerage, that will interact with the foreseen APIs of the blue data infrastructures. MARIS will interact with each Data & Access provider to check the functioning of their API which should allow the data brokerage service to retrieve data collections in sync with their agreed metadata entries, also taking into account the local data access mechanisms. One part is a shopping mechanism, directly linked to the Blue Cloud metadata catalogue, and individual adaptors dealing with the API's of Data & Access providers. The shopping mechanism will include AAA services for keeping track of requests and facilitating the delivery services to customers. Another part is a shopping ledger for users and Data & Access providers to keep track of shopping requests and progress by shopping adaptors for fulfilling the requests. The shopping components will be derived from the services that MARIS already has developed in the SeaDataCloud project as part of the SeaDataNet CDI service.

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

Integration of the Blue-Cloud data catalogue service and data broker will establish the Blue-Cloud data discovery and access service with GUI and API which is planned for launch end M18 (end March 2021). Overall, the steps from specification to development to integration to operational deployment to acceptance testing will be followed.