

Project Title	FAIR Earth Sciences & Environment services
Project Acronym	FAIR-EASE
Grant Agreement No.	101058785
Start Date of Project	01/09/2022
Duration of Project	36 Months
Project Website	fairease.eu

D2.4 - FAIR-EASE Data Discovery and Access Service - First Release

Work Package	WP2 - Discovery, Access and FAIR Data services
Lead Authors (Org)	Tjerk Krijger (MARIS), Peter Thijsse (MARIS)
Contributing Author(s) (Org)	Gwenaëlle Moncoiffé (NOC-BODC), Alexandra Kokkinaki (NOC-BODC), Alessandro Rizzo (IRD, Data Terra), Enrico Boldrini (CNR), Christelle Pierkot (CNRS, Data Terra)
Due Date	30.06.2024
Date	31.07.2024
Version	1.0

Dissemination Level

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



Versioning and contribution history

Version	Date	Author	Orcid ID	Notes
0.1	29.05.2024	Tjerk Krijger (MARIS)	0000-0002-1722-0523	Set-up structure of document
0.2	09.07.2024	Tjerk Krijger (MARIS)	0000-0002-1722-0523	Finalising the structure of the document
0.3	10.07.2024	Peter Thijsse (MARIS)	0000-0001-9214-3217	Review of content
0.4	17.07.2024	Tjerk Krijger (MARIS)	0000-0002-1722-0523	Draft version
0.5	19.07.2024	Alessandro Rizzo (IRD, Data Terra)	0000-0002-6085-648X	Review of content
0.6	24.07.2024	Alexandra Kokkinaki (NOC-BODC)	0000-0001-8042-6391	Added content to section 3.2
0.7	29.07.2024	Gwenaëlle Moncoiffé (NOC-BODC)	0000-0001-6559-4178	Review of content
0.8	30.07.2024	Peter Thijsse (MARIS)	0000-0001-9214-3217	Revision
1.0	31.07.2024	Christelle Pierkot (CNRS, Data Terra)	0000-0002-2591-3311	Final revision
Final				Final edition for submission

Disclaimer

This document contains information which is proprietary to the FAIR-EASE Consortium. Neither this document nor the information contained herein shall be used, duplicated or communicated by any means to a third party, in whole or parts, except with the prior consent of the FAIR-EASE Consortium.

Table of Contents

Contents

1. Introduction	7
1.1 Background	7
1.2 Objectives.....	7
2. Overview of data infrastructures integrated in IDDAS	9
2.1 First release of IDDAS	9
2.2 Data infrastructures to be integrated	9
3. Technical implementation	11
3.1 Outline of technical framework	11
3.2 FAIR-EASE DCAT model (DCAT-FE).....	12
3.2.1 DCAT-FE metadata	12
3.2.2 Mapping to DCAT-FE	15
3.2.3 DCAT-FE asset description example.....	17
3.2.4 Next steps for DCAT-FE	19
3.3 Search UI	21
3.4 SPARQL API.....	21
4. First release	22
4.1 Asset discovery.....	22
4.1.1 Search UI	22
4.1.2 SPARQL API	26
4.2 Data access.....	27
5. Integration of IDDAS	29
5.1 Integration with FAIR-EASE Earth Analytical Lab (EAL).....	29
5.2 Connection with Uniform Data Access Layer (UDAL)	30
6. Sustainability of results	31
7. Appendix	33

List of Figures

Figure 1 - Overall FAIR-EASE architecture.....	8
Figure 2 - Technical framework of the IDDAS.	11
Figure 3 - FE-DCAT class-diagram.....	13
Figure 4 - Mappings occurring in the FE DAB.....	15
Figure 5 - Users can enter text, such as keywords, titles, or descriptions and click on the button to perform their search based on the filters that they have provided.	22
Figure 6 - Time period filter.	22
Figure 7 - Geographic area filter.	22
Figure 8 - Parameter search list, users can enter letters to search through the list quickly.	23
Figure 9 - Catalogue filter, users can select the catalogue from a list.	23
Figure 10 - Media type filter, users can select the format they wish to retrieve.	24
Figure 11 - Link to copy the targeted SPARQL query to the user's clipboard.....	24
Figure 12 - Map interface using OpenStreetMap. The yellow box indicates the bounding box chosen by the user. The blue boxes are the resulting data sources found.	26
Figure 13 - SPARQL API input field.	27
Figure 14 - Results from a SPARQL query.	27
Figure 15 - Data access links to retrieve a data set found via the search UI.	28
Figure 16 - Access point of IDDAS from the FAIR-EASE EAL.....	29

Terminology

Terminology/Acronym	Description
API	Application Programming Language
CDI	Common Data Index
CF	Climate and Forecast
CMS	Content Management System
CSV	Comma-Separated Values
DAB	Discovery and Access Broker
DCAT	Data Catalog Vocabulary
DOI	Digital Object Identifier
EAL	Earth Analytical Lab
EOSC	European Open Science Cloud
FAIR	Findable; Accessible; Interoperable; Reusable
FE	FAIR-EASE
I-ADOPT	Interoperable Descriptions Of Observable Property Terminology
IDDAS	Interdisciplinary Data Discovery and Access Service
ISO	International Organization for Standardization
JSON	JavaScript Object Notation
NetCDF	Network Common Data Form
NVS	NERC vocabulary Server
TRL	Technology Readiness Level
RDF	Resource Description Framework
SA	Semantic Analyser
SPARQL	SPARQL Protocol and RDF Query Language
UDAL	Uniform Data Access Layer
UI	User Interface
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
VRE	Virtual Research Environment
WP	Work Package
XML	Extensible Markup Language
KER	Key Exploitable Result
DCMI	Dublin Core Metadata Initiative
WKT	Well-Known Text

Executive Summary

For researchers it is not easy to find, access and combine sets of multidisciplinary environmental data because the underlying data sources use their own data access mechanisms and (meta)data standards and formats, which makes it difficult and slow to collect and process the data needed for research. Therefore, we need to bridge the gaps between domain-specific (meta)data standards and provide the scientists with a harmonised way of finding, accessing and processing this (meta)data.

The FAIR-EASE (FE) Interdisciplinary Data Discovery and Access Service (IDDAS) tries to tackle these problems by creating a new standard metadata description, called the FAIR-EASE DCAT Model (DCAT-FE), to describe datasets from the different domains in the same metadata template. By doing so, FE WP2 has worked on a process that is more **bottom-up** and generic, making use of internationally accepted standards and the power of linked open data. This enables the IDDAS to provide harmonised search functionalities and access to various data sets from multidisciplinary environmental data sources. The IDDAS will make it possible to share and use data across different fields and connect with the FE Earth Analytic Lab (EAL) for analysis and visualisation.

This Deliverable describes the first release of the IDDAS, which is accessible for users and machines via the:

- **Search User Interface (UI)** - <https://fair-ease-iddas.maris.nl/search>: That lets researchers search across different assets using filters like time period, location, and data format;
- **SPARQL API** - <https://fair-ease-iddas.maris.nl/sparql>: That allows machines to search and access the assets.

It describes its background, objectives, and the data infrastructures it integrates. It further elaborates on the DCAT-FE described in D2.3 (<https://zenodo.org/records/10606930>), explaining here the mapping process for various types of data sources and highlighting future improvements. Furthermore, the IDDAS UI, SPARQL API, and methods for asset discovery and data access are explained and screenshots are provided. Ultimately the integration plans for the IDDAS with the FE EAL and the Uniform Data Access Layer (UDAL) are described, focussing also on the sustainability of the results.

1. Introduction

1.1 Background

Many of today's scientific problems are interconnected and involve multiple areas of knowledge. To solve these problems, it is important for scientists from different fields to work together on multidisciplinary case studies. For these studies to succeed, we need to bridge the gaps between different domain-specific (meta)data standards and provide the scientists with a harmonised way of finding, accessing and processing this varying (meta)data. The FAIR-EASE (FE) project is an example of such an initiative, which aims to build a technical and semantic framework to meet the needs of three different multidisciplinary Use Cases (UCs).

There is much overlap with thematics at EOSC level where there are similar challenges in providing data access to the large variety of datasets in the data infrastructures. While in the 2023 version of EOSC there was a top-down approach to make catalogues and their services available, FE WP2 has worked on a process that is more **bottom-up** and generic, making use of internationally accepted standards and the power of linked open data.

One part of the FE framework is the Interdisciplinary Data Discovery and Access Service (IDDAS) that helps users find and access various types of data sets (like e.g. in situ, satellite, model outputs) from existing environmental data sources. This service will make it possible to share and use data across different fields and connect with the FE Earth Analytic Lab (EAL) for analysis and visualisation. The IDDAS approach documents new recommended practices and demonstrates the added value.

1.2 Objectives

In virtual research environments, and other online data processing environments a returning challenge is to be able to find and access exactly the right data needed for further analysis, i.e. a subset of data with the targeted parameters, their units, and in a format appropriate for the processing software. That challenge is enormous, because each data infrastructure has its own metadata model, data format(s), metadata services, and data access services. The ultimate goal of the FE IDDAS is to develop a generically applicable solution to enable users to find and access (meta)data through an asset selector within the EAL (an example of a Virtual Research Environment in EOSC context).

This Deliverable, documents the **first release** of the FE IDDAS (not yet integrated in the EAL), which consists of a **search User Interface (UI)** and a **SPARQL API** for machine-to-machine access. It describes its background, objectives, and the data infrastructures it integrates. It further elaborates on the DCAT-FE described in D2.3 (<https://zenodo.org/records/10606930>), explaining here the mapping process for various types of data sources and highlighting future improvements. Furthermore, the IDDAS UI, SPARQL API, and methods for asset discovery and data access are explained and screenshots are provided. Ultimately the integration plans for the IDDAS with the FE EAL and the Uniform Data Access Layer (UDAL) are described, focussing also on the sustainability of the results including the expected Technology Readiness Levels (TRLs).

Figure 1 shows the location of the IDDAS within the complete FE architecture, including the connection to the EAL via the UDAL that will use the asset catalogue to discover relevant data collections and retrieve the corresponding assets.

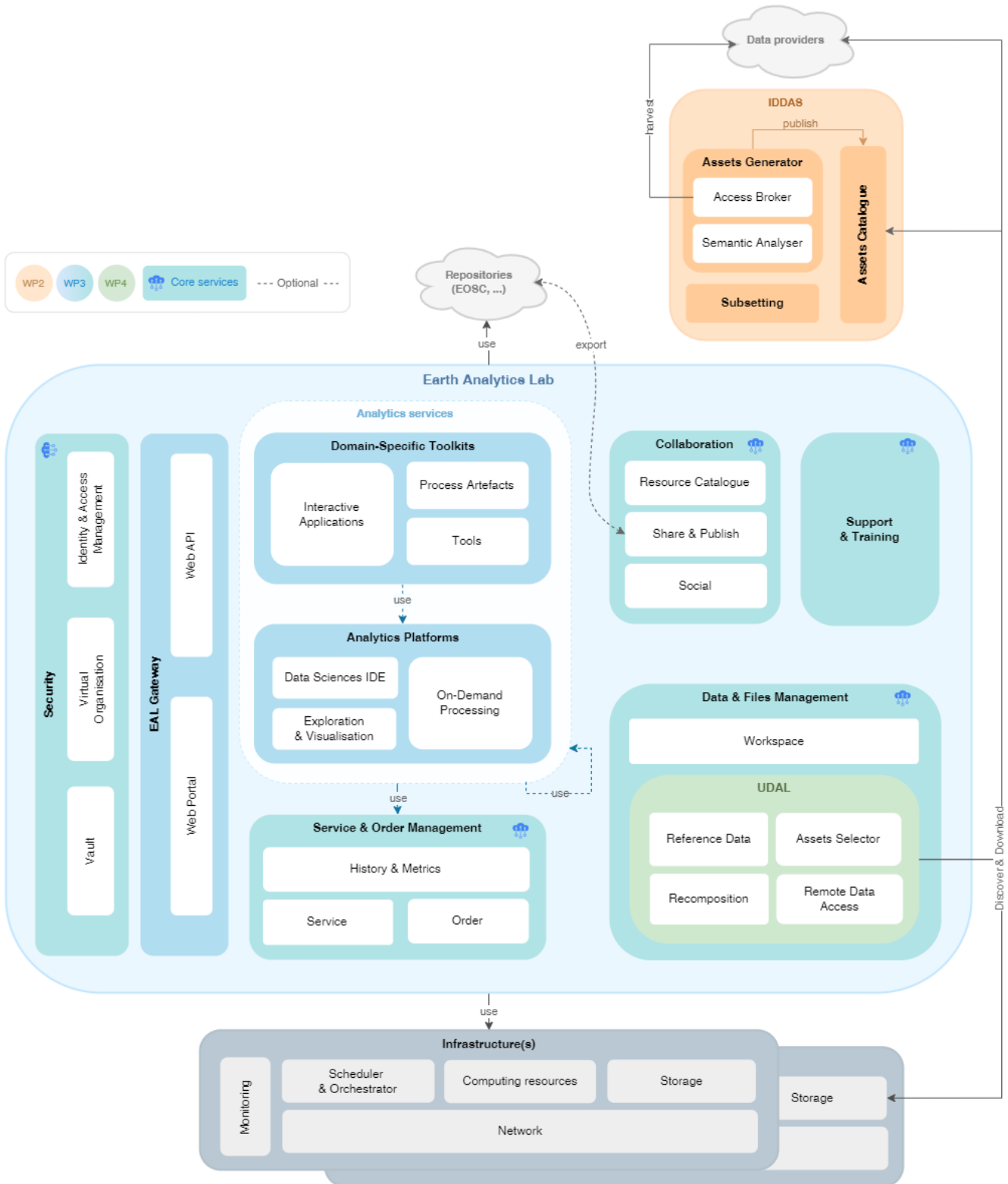


Figure 1 - Overall FAIR-EASE architecture.

2. Overview of data infrastructures integrated in IDDAS

2.1 First release of IDDAS

The first release of the FE IDDAS includes a UI and SPARQL API that allows users to search for (meta)data assets from multi-disciplinary data infrastructures that were identified at the start of the project by the pilots as being essential for conducting their research. The infrastructures that are included in the first version of the IDDAS are:

- VITO / Copernicus Global Land Services
 - <https://land.copernicus.eu/en>
- Copernicus Marine Environment Monitoring Service (CMEMS)
 - <https://data.marine.copernicus.eu/products>
- WEKEO
 - <https://www.wekeo.eu/>
- US NODC Collections
 - <https://www.ncei.noaa.gov/>
- Joint Research Centre Data Catalogue
 - <https://data.jrc.ec.europa.eu/>
- European Environment Agency SDI Catalog
 - <https://www.eea.europa.eu/en>
- EMODnet Bathymetry
 - <https://emodnet.ec.europa.eu/en/bathymetry>
- Euro-Argo
 - <https://www.euro-argo.eu/>
- ELIXIR-ENA
 - <https://www.ebi.ac.uk/ena/browser/home>
- EMODnet Chemistry
 - <https://emodnet.ec.europa.eu/en/chemistry>
- EMSO
 - <https://emso.eu/>
- EMODnet Biology
 - <https://emodnet.ec.europa.eu/en/biology>
- ICOS data portal
 - <https://www.icos-cp.eu/data-services/about-data-portal>
- SIOS
 - <https://sios-svalbard.org/sios-ri-catalogue>
- SeaDataNet products
 - <https://www.seadatanet.org/Products#/search?from=1&to=30>

2.2 Data infrastructures to be integrated

There are also a set of infrastructures listed below which have yet to be integrated. Some of these infrastructures can be included via the same approach as the ones listed above (sections 3.2.2.1 & 3.2.2.2 for more information), while some others will need a different approach (section 3.2.2.3).

- NASA Earth Observation Data
 - <https://www.earthdata.nasa.gov/>
- Copernicus Data Space Ecosystem
 - <https://dataspace.copernicus.eu/>
- EUMETSAT
 - <https://www.eumetsat.int/>
- Climate Data Store
 - <https://cds.climate.copernicus.eu/cdsapp#!/search>
- Physical Sciences Laboratory
 - <https://psl.noaa.gov/>
- European Soil Data Centre
 - <https://esdac.jrc.ec.europa.eu/>
- National Center for Biotechnology Information (NCBI)
 - <https://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/index.cgi>
- Global Biodiversity Information Facility (GBIF)
 - <https://www.gbif.org/>
- World Register of Marine Species
 - <https://www.marinespecies.org/>
- European Marine Omics Biodiversity Observation Network
 - <https://www.embrc.eu/services/emo-bon>
- SEANOE
 - <https://www.seanoe.org/>
- SeaDataNet
 - <https://cdi.seadatanet.org/search>
- EcoTaxa
 - <https://ecotaxa.obs-vlfr.fr/>
- SOCAT
 - <https://socat.info/>

3. Technical implementation

3.1 Outline of technical framework

The IDDAS UI and SPARQL API are designed for discovery and access to multi-disciplinary datasets and are built on top of a catalogue of (meta)data assets that are harvested from the data infrastructures listed in chapter 2 (see also Figure 2). The solution for the IDDAS is based on a **bottom-up** description of the “data assets” that are available via their data access service supporting file-based systems as well as subsetting services. This description is given in a custom-made FE DCAT model (DCAT-FE) that describes these multi-disciplinary assets uniformly. This allows users to find and choose from a wide range of datasets in a harmonised manner and search for relevant data using a faceted search based on keywords connected to the metadata and depending on the metadata-completeness. Furthermore, this asset description FAIR bottom-up approach was important because it allowed for involving the end-user in the creation of the DCAT-FE metadata elements, making sure that it fits the needs and enables sufficient querying possibilities.

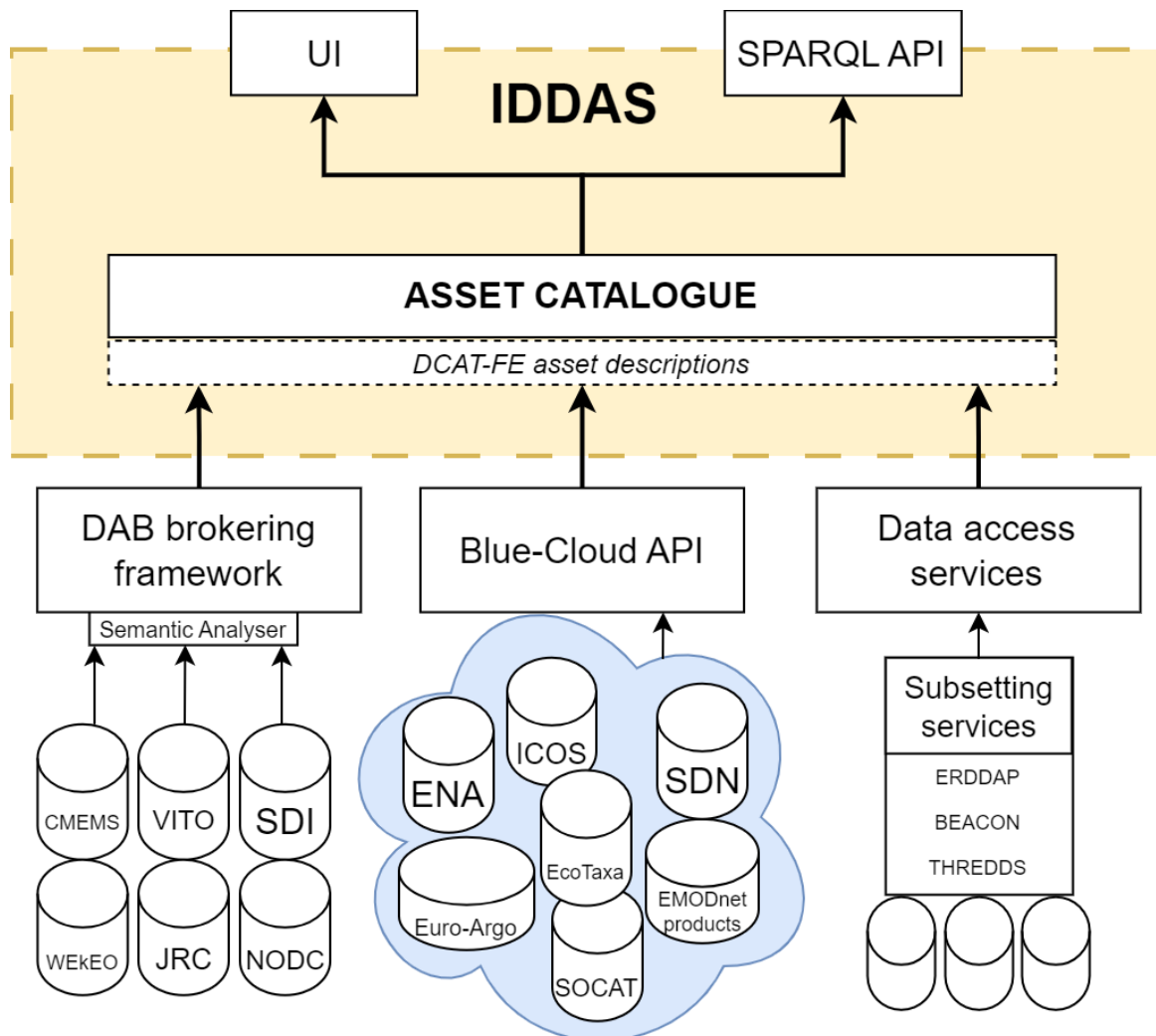


Figure 2 - Technical framework of the IDDAS.

The development of the IDDAS involves several key components, which are illustrated in Figure 2. The Figure illustrates three different harvesting approaches for obtaining DCAT-FE assets from the data infrastructures:

- Blue-Cloud API: These assets are exposed via the Blue-Cloud API and concern Euro-Argo ELIXIR-ENA, EMODnet Chemistry, EMSO, EMODnet Biology, ICOS data portal, SIOS and SeaDataNet products. Section 3.2.2.2 describes the mapping process.
- DAB brokering framework: These assets are exposed via the DAB broker and concern VITO/Copernicus Global Land Services, Copernicus Marine Environment Monitoring Service (CMEMS), WEKEO, US NODC catalog, Joint Research Centre Data catalog, European Environment Agency SDI Catalog and EMODnet Bathymetry. Section 3.2.2.1 describes the mapping process.
- Data access services: The assets that are exposed via data access services like an FTP, ERDAPP or Beacon can often lack sufficient metadata for mapping it towards DCAT-FE. In those cases, data access service itself will be described in DCAT-FE and published in the asset catalogue, making the service findable and accessible rather than its outputs.

The assets that are included in the first release of the IDDAS are created via the Blue-Cloud API and the DAB brokering framework. In the next iteration (September), DCAT-FE assets will also be created for the data access services.

The next sections will dive into the DCAT-FE asset descriptions in more detail and will describe the technical implementation of the UI and SPARQL API on top of the asset catalogue that includes the DCAT-FE assets.

3.2 FAIR-EASE DCAT model (DCAT-FE)

The DCAT-FE model is a specification based on the Data Catalogue vocabulary (DCAT 3.0¹), for describing cross-disciplinary datasets originating from various research infrastructures in Europe and utilised by the FE project pilots (WP5). It is an RDF vocabulary designed to facilitate interoperability between data catalogues. The uniform asset description of DCAT-FE serves the following purposes:

- Harmonise the common metadata identified from the underlying data sources;
- Enable semantic searches: IDDAS utilises this information to facilitate semantic searches of the underlying datasets;
- Facilitate aggregation and harmonisation: the described assets are used to support the aggregation and harmonisation of data provided via the project's UDAL.

The assets are harvested from, or manually created for, the data infrastructures. They are then published using a SPARQL endpoint on top of which the UI is built.

3.2.1 DCAT-FE metadata

The DCAT-FE development started during development cycle 2 and was updated during the third and fourth cycle. The objective of the DCAT-FE profile is to describe all available datasets in a standardised way, while aligning with the application requirements of the FE project. DCAT-FE is

¹ <https://www.w3.org/TR/vocab-dcat-3/>

The metadata elements that are covered in DCAT-FE are listed in Table 1. The quality and completeness of their content will depend on the quality and completeness of metadata input from the data infrastructure at the source. The first release of the IDDAS that is documented here is built on top of this first version of the DCAT-FE. However, the DCAT-FE profile has the flexibility to allow for future improvements. The use of DCAT-FE enables us to describe heterogeneous data sources in a standardised way with semantic annotation, while keeping the source ones intact. This enables the IDDAS UI and the SPARQL API (and ultimately the asset selector in the EAL) to access a semantically and syntactically harmonised asset catalogue and the UDAL to aggregate such descriptions into a single point of access. Ultimately, it allows the pilot researchers to easily find distributed datasets from one central location.

Table 1 - The metadata elements currently covered by DCAT-FE.

Term	Description
Top level parameter	Broad level parameters (e.g. P05, P22, etc.)
Parameter	Measured variables (e.g. P01, P02, GCMD, etc.)
Quality flag	Flags used to provide information on data quality or data values
Time period	Start-end time of the dataset
Bounding box / Polygon, i.e. spatial region	Spatial description of the region of interest as a bounding box or a polygon
Download URL	Link directly allowing for downloading of the dataset
Data format	Format of the dataset (e.g. NetCDF, CSV, ODV, etc.)
Access services	Describes the protocol for accessing the dataset (e.g. direct download, WMS, WFS, etc.)
Data collection/data infrastructure	The data infrastructure the dataset is made available from
Unique identifier	A unique identifier to refer to the dataset (e.g. CDI, DOI)
Platform code	The code of the platform that the sensor was attached to when the measurement took place (e.g. WMO, C17, ICES)
Instrument type	The type of instrument used for the measurement (e.g. L22, L05)
Coordinate system	Coordinate reference system of the dataset (e.g. L10)
Title	Title of the dataset
Abstract/description	A short description of the context / content of the dataset
Publisher	The organisation/institution who published the dataset (e.g. EDMO, ROR)
Creator	The organisation/institution who created the resource (e.g. EDMO, ROR)

3.2.2 Mapping to DCAT-FE

This section explains how the assets from the multi-disciplinary data infrastructures are exposed in DCAT-FE using three different harvesting approaches: 1) DAB brokering framework; 2) Blue-Cloud API and 3) Data access services.

3.2.2.1 DAB brokering framework

The FE DAB component is a brokering framework powered by the Discovery and Access Broker (DAB) technology. It is deployed in the context of FE to perform syntactic harmonisation of the heterogeneous metadata documents made available by the different sources and enabling on top of them uniform discovery capabilities. The DAB harmonised model is based on the ISO-19115 metadata model (having ISO 19139 as its XML schema encoding), comprising a predefined list of more than 400 metadata elements and supporting as well custom extensions, useful to harmonise possible metadata elements that are outside of ISO-19115 scope.

The diagram in Figure 4 shows the mapping procedure occurring in the FE DAB and explains the following:

1. The information flow is started by the FE data providers that publish online metadata describing the available data resources. Each of them is in general described according to a specific data model and encoded in a specific machine-readable metadata format (e.g. EML, RDFS, CDI, etc.).
2. A first metadata mapping takes place starting from the original metadata harvested by the FE broker from each FE provider. Each metadata element is syntactically translated to correspondent elements in the DAB harmonised metadata model, based on ISO-19115 and supporting custom community extensions.
3. A second mapping takes place starting from the harmonised metadata model, and has the target DCAT-FE model as the output as required by the FE asset catalogue.

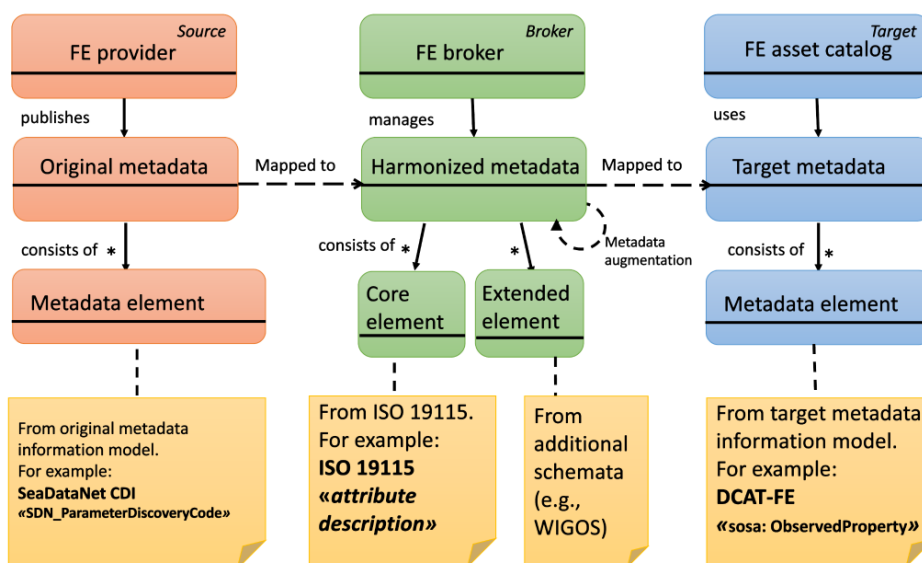


Figure 4 - Mappings occurring in the FE DAB.

Most of the DCAT-FE elements can be mapped from correspondent ISO-19115 metadata elements. For the ones that don't yet have a place in the baseline ISO-19115 (e.g. taxonomic cover, centroids,

schema column information) custom extensions can be made based on their definitions from DCAT. In some cases, it is possible to leverage the keywords section as a built-in extension mechanism, as this is a metadata section already supporting codes from controlled vocabularies and vocabulary descriptions. Multiple keyword types can be managed, by using appropriate keyword type codes. Regarding this specific point, the possible keyword type codes in the last version of ISO-19115 specification has increased to include: *discipline, place, stratum, temporal, theme, dataCenter, featureType, instrument, platform, process, project, service, product, subTopicCategory* and *taxon*. This is what the Semantic Analyser (SA) can help doing in the future, i.e. place a "keyword" in its most probable "type code" category. Additional keyword type codes could be defined as well, according to the FE requirements.

3.2.2.2 Blue-Cloud API

The Blue-Cloud Data Discovery and Access Service (DDAS) currently provides access to eleven Blue Data Infrastructures (BDIs). Each of these BDIs contain datasets at Collection Level (level 1) and Granule Level (level 2):

1. Collection Level: Discovery of data collections using a metadata profile based on ISO-19115, with focus on aggregated collection-level data through free search, geographic search, and temporal filters. This metadata profile is the same for each BDI.
2. Granule/Data Level: Detailed discovery within collections to access specific data granules via heterogeneous web services (e.g. OGC-CSW, ERDDAP, APIs), using additional search criteria specific to each data infrastructure. The associated metadata differs between the different BDIs, while some BDIs are equal for level 1 and level 2.

In the context of FE, we have access to the Blue-Cloud API, which gives us machine-to-machine access to the (meta)data records from the eleven BDIs that are currently in the system. The approach for generating DCAT-FE assets from the Blue-Cloud API is as follows:

- For each BDI record, the metadata from level 1 and level 2 are combined into one larger metadata file, in order to provide all available metadata. E.g. for a Euro-Argo record the generic level 1 terms such as title, organisation, bounding box are combined with (specific to Euro-Argo) level 2 terms like cycle id, platform name, etc.
- The combined metadata file for each record is thereafter mapped towards DCAT-FE, following a mapping exercise on the metadata elements.
- With the mapping now available, the metadata records are harvested from Blue-Cloud and combined in a triple database and published via a SPARQL endpoint.

It is important that the assets from these BDIs are going through Blue-Cloud, since this is the organised European data infrastructure for marine access. Blue-Cloud already works on harmonising data and services and is looking to become an EOSC node. When improvements are made on the Blue-Cloud side on the metadata model and metadata completeness, this will thus also immediately be implemented on the FE side via the Blue-Cloud API. The other way around it can also be the case that when FE introduces new semantic improvements, this can be implemented in Blue-Cloud in the next iterations.

3.2.2.3 Generating assets from data access services

A selection of the data infrastructures does not allow for harvesting of the (meta)data by broker services such as the DAB, but rather they offer data access services (e.g., FTP, ERDDAP, THREDDS

servers, etc.) often with inconsistent metadata (a lack of metadata standardisation and/or minimal metadata). In those cases, we do two things:

- Give recommendations to the data infrastructures for improving their FAIRness. This can be done by providing guidelines and best practices to the data infrastructures on improving their metadata standards, quality and coverage. But also providing them with information on well-known services (WMS, WFS, etc.) they can implement that allow for harvesting of their metadata.
- Describe the data access services that are available at the data infrastructures in DCAT-FE and publish them in the asset catalogue, making the service findable and accessible rather than its outputs. In DCAT-FE, the service should therefore, be well-described along with the protocol for data access (section 3.2.4.2). It should be clear for a human, as well as a machine, how the access service can be queried. This way, a user can find such a service using the UI or SPARQL API, by querying for example on: “access protocol FTP, parameter Oxygen” and access the service with targeted queries.

For data infrastructures with such data access services that require sub-setting by the pilots, a sub-setting service can be implemented on top of those infrastructures. This specific instance of the sub-setting service is then also described in DCAT-FE and published in the asset catalogue. In essence, this would be described in the same way as the data access services mentioned above. Only, the data service description in DCAT-FE will be different.

By implementing sub-setting services such as for example Beacon³, we gain control over the output, enabling us to present also the output assets in the proposed DCAT-FE standard. This way we can publish a set of often-used subsets in the asset catalogue. For example, by publishing multiple subsets based on common regions, parameters, time periods and depth ranges. In this way a whole set of DCAT-FE assets will be created. If the user requires a more dynamic query approach, by creating its own request, the Beacon system itself can be used and accessed from the IDDAS.

3.2.3 DCAT-FE asset description example

As an example for an asset description in DCAT-FE, we will take a look at a dataset from the European Environment Agency SDI catalogue, called “O3, European air quality data for 2015, (interpolated data)”. The data set provides concentrations for the air pollutants Ozone (O3) at 1 km grid. The dataset can be found in the IDDAS:

- <https://fair-ease-iddas.maris.nl/dataset?uri=https://essi-lab.eu/dab/fair-ease/dataset/ead3f70c-ba7e-4c2c-8073-f79a52bf1aae>

For each part of the RDF data snippet there is an explanation given on what it is indicated.

```
@prefix schema: <https://schema.org/> .
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix dc: <http://purl.org/dc/terms/> .
@prefix geo: <http://www.opengis.net/ont/geosparql#> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
```

³ Beacon is a data lake solution used to store and subset millions of NetCDF datasets and terabytes of data. For more information see <https://beacon.maris.nl/>.

Prefixes help in distinguishing between terms from different vocabularies that might have the same local name. For instance, `dc:dataset` from the Dublin Core namespace can be distinguished from `schema:dataset` from the `schema.org` namespace. The prefix schemas used here, are:

- The namespace <https://schema.org/>, which provides a vocabulary for describing structured (meta)data;
- The namespace <http://www.w3.org/ns/dcat#>, which is used for describing datasets and data catalogues;
- The namespace <http://purl.org/dc/terms/>, which is the Dublin Core Metadata Initiative (DCMI) terms namespace, used for general metadata terms;
- The namespace <http://www.opengis.net/ont/geosparql#>, which is a standard for representing and querying geospatial data on the semantic web;
- The namespace <http://www.w3.org/2001/XMLSchema#>, which is the XML schema definition namespace, used for defining data types.

<https://essi-lab.eu/dab/fair-ease/dataset/ead3f70c-ba7e-4c2c-8073-f79a52bf1aae>

This is the URI of the dataset being described.

`a schema:Dataset, dcat:Dataset ;`

This states that the subject is of types `schema:Dataset` and `dcat:Dataset`, indicating it is a dataset according to both `schema.org` and `DCAT` vocabularies.

`dc:description "The data set provides concentrations for the air pollutants Ozone (O3) at 1 km grid combining monitoring air quality data in a 'regression-interpolation-merging mapping' methodology and the observational values of the air quality monitoring stations used in the interpolation. It covers Europe for the year 2015. It provides estimates for human health related indicators of pollutants ozone (93.2 percentile of maximum daily 8-hour means, SOMO35) and vegetation related ozone indicators (AOT40 for vegetation and for forests) for the year 2015." ;`

This provides a description of the dataset.

`dc:identifier "eea_r_3035_1_km_aq-interpolated-03_p_2015_v01_r00" ;`

This gives a unique identifier for the dataset.

`dc:issued "" ;`

This could provide the issuance date of the dataset, but it is empty here.

```
dc:spatial [
  a dc:Location ;
  dcat:bbox ""POLYGON((
    -25.0 34.0, 45.0 34.0, 45.0 72.0, -25.0 72.0, -25.0 34.0))""^^geo:wktLiteral
];
```

This describes the spatial coverage of the dataset. It indicates that the dataset covers a polygon defined in Well-Known Text (WKT) format, representing geographic coordinates.

```
dc:temporal [
  a dc:PeriodOfTime ;
  dcat:endDate "2015-12-31"^^xsd:date ;
  dcat:startDate "2015-01-01"^^xsd:date
];
```

This describes the temporal coverage of the dataset, specifying the start and end dates of the data period.

`dc:title "O3, European air quality data for 2015, (interpolated data)" ;`

This gives the title of the dataset.

```
dcats:title <https://essi-lab.eu/dab/fair-ease/dataset/ead3f70c-ba7e-4c2c-8073-f79a52bf1aae/distribution/1>, <https://essi-lab.eu/dab/fair-ease/dataset/ead3f70c-ba7e-4c2c-8073-f79a52bf1aae/distribution/0> ;
```

This lists the distributions (access points) for the dataset provided by FAIR-EASE.

```
dcats:theme "air quality monitoring", "transboundary pollution", "atmospheric composition", "European", "Measurement and modelling data (Air Quality Directive)", "air quality", "troposphere", "Air pollution", "Monitoring stations (Air Quality Directive)", "atmospheric particulate", "traffic", "atmospheric pollution", "ozone", "Directive 2008/50/EC", <http://inspire.ec.europa.eu/theme/ef>, <http://inspire.ec.europa.eu/theme/ac>, <http://inspire.ec.europa.eu/theme/hh> .
```

This specifies themes related to the dataset using both literal strings (if URIs are not available) and URIs from the INSPIRE directive, covering different aspects such as air quality, atmospheric composition, and pollution. An evolution will be done with the Semantic Analyser to implement URIs where possible and better comply with the FAIR principles (section 3.2.4.1).

```
<https://essi-lab.eu/dab/fair-ease/dataset/ead3f70c-ba7e-4c2c-8073-f79a52bf1aae/distribution/0>
```

This is the URI of the first distribution of the dataset.

```
a dcat:Distribution ;
```

This states that the subject is of type dcat:Distribution.

```
dcats:accessService <https://essi-lab.eu/dab/fair-easesubset-service--2014000675> ;
```

This points to a service providing access to the distribution.

```
dcats:accessURL <https://sdi.eea.europa.eu/webdav/datastore/public/eea_r_3035_1_km_aq-interpolated-03_p_2015_v01_r00/> .
```

This gives the URL where the distribution can be accessed.

```
<https://essi-lab.eu/dab/fair-ease/dataset/ead3f70c-ba7e-4c2c-8073-f79a52bf1aae/distribution/1>
```

This is the URI of the second distribution of the dataset.

```
a dcat:Distribution ;
```

This states that the subject is of type dcat:Distribution.

```
dc:title "Direct download" ;
```

This provides a title for the distribution.

```
dcats:accessService <https://essi-lab.eu/dab/fair-easesubset-service--2014000675> ;
```

This points to a service providing access to the distribution.

```
dcats:accessURL <https://sdi.eea.europa.eu/data/ead3f70c-ba7e-4c2c-8073-f79a52bf1aae> .
```

This gives the URL where the distribution can be accessed.

3.2.4 Next steps for DCAT-FE

3.2.4.1 Semantic annotation

In order to semantically harmonise the converted datasets according to the DCAT-FE we need to be able to unambiguously identify the semantic assets used by the data infrastructures for describing

key metadata elements, focusing on parameters, instruments, and platforms. This step is greatly facilitated if the research infrastructures annotate their data assets using both, terms from controlled vocabularies and their associated Uniform Resource Identifiers (URIs). However, this is not currently a well-established practice. For this reason, we have developed a tool, the SA, that allows us to:

- Identify terms used by the different RIs to describe parameters, instruments, and platforms;
- Find possible matches with known semantic resources held in our FE Knowledge Base;
- Obtain the unique URI for the terms used if these were missing;
- Offer possible mappings of the terms used to those originating from the reference list of vocabularies identified in D2.3⁴.

The SA was described in detail in D2.3. New functionality has since been added, in particular the ability to submit lists of terms in a text file. This is a useful addition for users seeking to evaluate a list of terms against known semantic resources. We have also added the option to semantically analyse GeoDAB reports that are produced for the Blue-Cloud research infrastructures as part of the Blue-Cloud2026 project. This gives us a more synoptic analysis of all the datasets offered by a research infrastructure, as opposed to performing single dataset analysis one at a time. A new search box was also added to enable users to target specific datasets based on a simple keyword search.

We are continuing to improve the tool, refine its underlying matching methods and Knowledge Base to obtain more targeted results. The next steps will be to provide outputs that could be used by the FE DAB to refine its semantic interpretation of the incoming data, and evaluate how the SA can be optimally used to semantically enhance the IDDAS and improve semantic searches.

3.2.4.2 Protocols for data access

As part of the FE asset catalogue, we will compile a list of data access protocols with standardised descriptions that will not only help developers navigate the API documentation and software libraries, but also enable machines (generic clients) to effectively select and execute the targeted data-requests. In a subsequent phase, this process might be further automated so that a generic request from the EAL can select an appropriate data access protocol along with instructions on how to access it, resulting in a list of specific subset URLs for seamless retrieval and further processing. The DCAT-FE will hence, need to be extended with protocols for data access (e.g., WMS, direct download links, WFS, etc.), which will first be human-readable and later machine to machine:

1. We will establish a list of Uniform Resource Identifiers (URIs) for each data access protocol that is provided by the data infrastructures.
 - a. This list will be hosted and managed through the FE GitHub repository:
 - i. <https://github.com/fair-ease/dataset-demand-register>
 - b. The domain for these URIs is <https://lab.fairease.eu/>. This domain will serve as the base URL for accessing and referencing the protocols.
 - c. For the DAB, the protocols are extracted here:
 - i. <https://blue-cloud.geodab.eu/gs-service/metadata-report/distribution-report.jsp?view=fair-ease>
2. These URIs will then be integrated into the DCAT-FE. Specifically, they will be included under the "conforms to" property within the metadata description. This integration ensures that

⁴ <https://zenodo.org/records/10606930>

each dataset within the DCAT-FE references the protocol(s) that are available. By incorporating URIs directly into the metadata, we establish a standardised approach for documenting and accessing protocols.

3.2.4.3 Subsets

DCAT-FE needs to accommodate details about subsets of original data files, in order to help users search the catalogue and choose exactly what they need, like smaller parts of big files, parts of several files, or groups of files, depending on their needs and what the service can handle. This will be included in the “distribution” of the DCAT-FE, which describes the type of access service. There, the query or key/identifier can be included that is used for querying the sub-setting service. This can then be used by the UDAL/client to go to the sub-setting service and retrieve the sub-set.

3.3 Search UI

The search UI of the IDDAS is available at <https://fair-ease-iddas.maris.nl/search> and is built on top of the SPARQL endpoint. The SPARQL is the only database behind the UI, which makes it so that all human searches can be translated to a machine-to-machine query that can be used directly on the SPARQL endpoint.

3.4 SPARQL API

The SPARQL API is accessible at <https://fair-ease-iddas.maris.nl/sparql> and currently contains the assets provided via the DAB brokering framework and the Blue-Cloud API. In order to set-up the SPARQL endpoint the following steps are followed:

1. The Blue-Cloud records are published in RDF (DCAT-FE) and harvested and collected in one big triple file.
2. The DAB records are published as RDF (DCAT-FE) and harvested with a SPARQL endpoint.
3. The Blue-Cloud and DAB records are combined and loaded into one big triple database (Apache DTB).
4. This database is published via a SPARQL endpoint - <https://fair-ease-iddas.maris.nl/sparql>

4. First release

4.1 Asset discovery

In this section we will discuss how assets can be found and describe the main search features of the **IDDAS search UI** and **SPARQL API**, including screenshots and some explanations. In section 4.2, we will discuss the access to data after the user has found their relevant sources.

4.1.1 Search UI

The search UI (<https://fair-ease-iddas.maris.nl/search>) is designed to be user-friendly, making it easy for users to search and access data using a set of common filters. And because it is built on an RDF database, the current set of filters can easily be expanded too. The search results that are displayed depend on the metadata completeness of the provided data sources. The more complete the metadata of a source, the higher the chance that the result will be displayed. In this first release of the IDDAS, not all data sources available at the data infrastructures that are listed in chapter 2 are included yet, which is why there are currently 30,000+ assets findable.

4.1.1.1 Free text search

The free text search allows users to find relevant data based on the title, description, and keywords available through the DCAT-FE.





Figure 5 - Users can enter text, such as keywords, titles, or descriptions and click on the button to perform their search based on the filters that they have provided.

4.1.1.2 Time period filter

Users can narrow down their search results by specifying a time period, using a start date and end date.



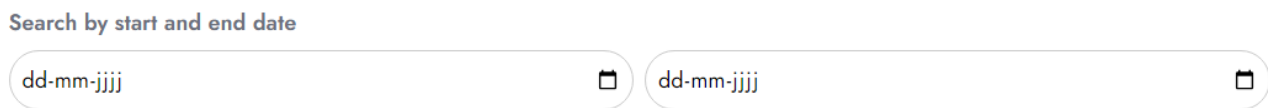


Figure 6 - Time period filter.

4.1.1.3 Geographic area filter

This feature allows users to filter data based on a geographic area, by selecting a bounding box with minimum and maximum values for longitude and latitude.



Figure 7 - Geographic area filter.

4.1.1.4 Parameter search

The search interface allows the users to search on parameters, by choosing from a parameter list. This list is directly extracted from the DCAT-FE metadata and does not currently include mapping of vocabularies, which is why for some parameters there are many different options available. This semantic mapping is foreseen to be done with the SA, which was discussed in more detail in D2.3. The idea is that there will be an API available, to which keywords/strings can be sent, for example P_sal (Practical salinity). The API would then return a mapping to for example (a set of) P01⁵, or R03⁶ terms, which could then be used for searching.

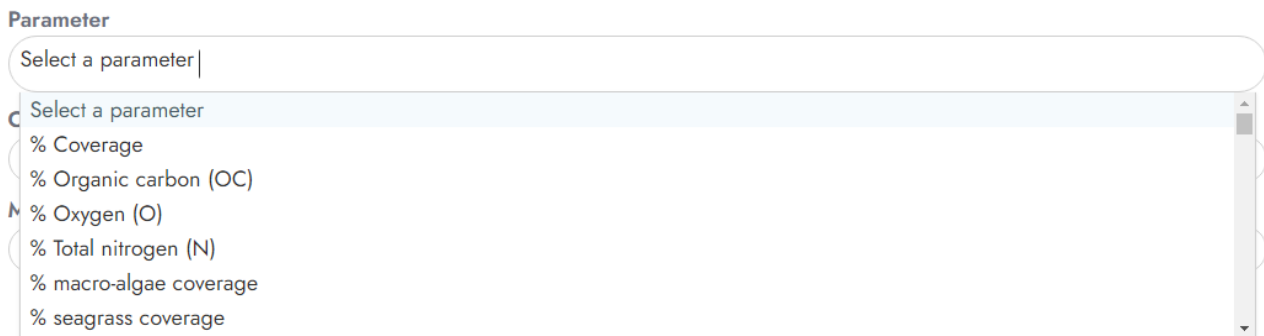


Figure 8 - Parameter search list, users can enter letters to search through the list quickly.

4.1.1.5 Catalogue filter

This filter allows users to specify the data infrastructures from which the data is harvested. The list currently includes the data infrastructures; VITO / Copernicus Global Land Services, European Environment Agency SDI Catalog, WEKEO, ELIXIR-ENA, Euro-Argo, EMODnet Biology, EMODnet Bathymetry, EMODnet Chemistry, ICOS data portal, SeaDataNet products, Copernicus Marine Environment Monitoring Service (CMEMS), US NODC Collections, Joint Research Centre Data Catalogue, SIOS and EMSO.

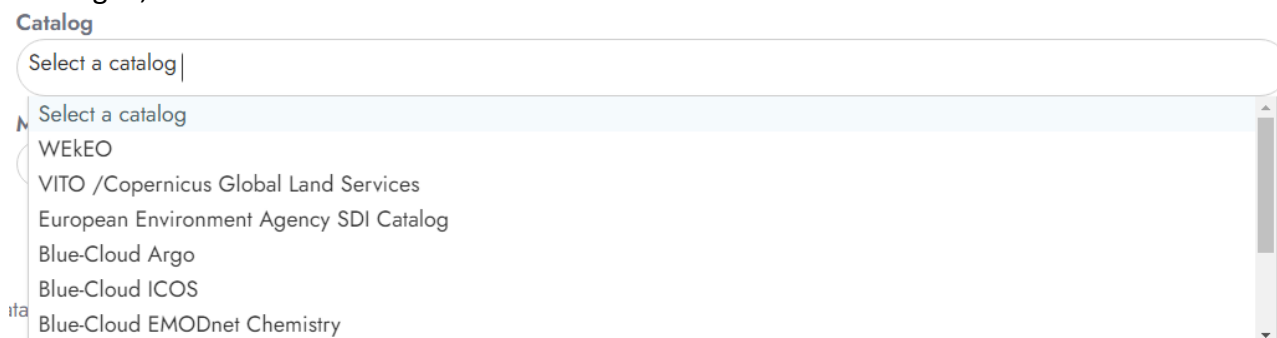


Figure 9 - Catalogue filter, users can select the catalogue from a list.

4.1.1.6 Media type (format) filter

This feature lets users filter data sets based on the media type (format) such as CSV, JSON, XML, etc. Currently it only contains the option NetCDF.

⁵ <https://vocab.nerc.ac.uk/collection/P01/current/>

⁶ <https://vocab.nerc.ac.uk/collection/R03/current/>

Media type

Select a media type |

Select a media type

NetCDF

ZIP

Figure 10 - Media type filter, users can select the format they wish to retrieve.

4.1.1.7 SPARQL query copy feature

After the user has selected the filters, they want to apply, there is an option to copy a targeted SPARQL query to their clipboard for those specific selection criteria. For example, when we search via free text “Air Quality” and via the catalogue filter “European Environment Agency SDI Catalog”, we obtain 180 data sets, with a SPARQL query shown below. At the bottom of the SPARQL query you can see how the filters are included.

Results

Found 180 datasets. [Copy SPARQL query for this page.](#)

Figure 11 - Link to copy the targeted SPARQL query to the user’s clipboard.

```
PREFIX dcat: <http://www.w3.org/ns/dcat#>
PREFIX dc: <http://purl.org/dc/terms/>
PREFIX prov: <http://www.w3.org/ns/prov#>
PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
PREFIX geo: <http://www.opengis.net/ont/geosparql#>
PREFIX schema: <https://schema.org/>
```

The prefix schemas used, stand for:

- The namespace <http://www.w3.org/ns/dcat#>, which is used for describing datasets and data catalogues;
- The namespace <http://purl.org/dc/terms/>, which is the Dublin Core Metadata Initiative (DCMI) terms namespace, used for general metadata terms;
- The namespace <http://www.w3.org/ns/prov#>, which is used for describing provenance information;
- The namespace <http://www.w3.org/2001/XMLSchema#>, which is the XML schema definition namespace, used for defining data types;
- The namespace <http://www.opengis.net/def/function/geosparql/>, which is used for geospatial functions in GeoSPARQL;
- The namespace <http://www.opengis.net/ont/geosparql#>, which is a standard for representing and querying geospatial data on the semantic web;
- The namespace <https://schema.org/>, which provides a vocabulary for describing structured (meta)data.

```
SELECT DISTINCT (GROUP_CONCAT(DISTINCT ?_title; SEPARATOR=' ') AS ?title) (GROUP_CONCAT(DISTINCT
?_mediaType; SEPARATOR='|') AS ?mediaType) ?dataset ?description ?bbox (REPLACE(STR(?catalog),
'\\?.*$', '' ) AS ?cleanCatalog)
```

This indicates the variables to be returned by the query. It concatenates distinct titles and media types, retrieves the dataset, description, bounding box, and a cleaned catalogue URI.

WHERE {

The WHERE clause specifies the pattern to match in the data.


```
?dataset a dcat:Dataset ;
  dc:title ?_title ;
  dc:description ?description .
```

This pattern matches datasets with a title and description.

```
OPTIONAL {
  ?dataset dc:temporal [
    a dc:PeriodOfTime ;
    dcat:startDate ?startDate ;
    dcat:endDate ?endDate
  ] .
}
```

This optional block retrieves the temporal coverage (start and end dates) of the dataset if available.

```
OPTIONAL {
  ?dataset dc:spatial [
    a dc:Location ;
    dcat:bbox ?bbox
  ] .
}
```

This optional block retrieves the spatial coverage (bounding box) of the dataset if available.

```
OPTIONAL {
  ?dataset schema:variableMeasured [
    a schema:PropertyValue ;
    schema:name ?parameterName
  ] .
}
```

This optional block retrieves any measured variables for the dataset if available.

```
OPTIONAL {
  ?dataset prov:used ?used .
}
```

This optional block retrieves any provenance information for the dataset if available.

```
OPTIONAL {
  ?catalog a dcat:Catalog ;
  dcat:dataset ?dataset .
}
```

This optional block retrieves the catalogue to which the dataset belongs if available.

```
OPTIONAL {
  ?dataset dcat:distribution ?distribution .
  ?distribution a dcat:Distribution ;
  dcat:mediaType ?_mediaType .
}
```

This optional block retrieves the distribution and its media type for the dataset if available.

```
FILTER(REGEX(?_title, '(Air|quality)', 'i') || REGEX(?description, '(Air|quality)', 'i') ||
REGEX(?used, '(Air|quality)', 'i')) .
```

This filter ensures that only datasets with titles, descriptions, or used entities containing "Air" or "quality" (case-insensitive) are included.

```
FILTER (BOUND(?catalog) && STRSTARTS(STR(?catalog), 'https://essi-lab.eu/dab/fair-
ease/catalog/UUID-456602db-4275-4410-8b68-436fd23ace69') ) .
```

This filter ensures that only datasets belonging to a specific catalogue URI are included.

```
} GROUP BY ?dataset ?description ?startDate ?endDate ?bbox ?used ?parameterName ?catalog LIMIT 10
OFFSET 0
```

This groups the results by the specified variables and limits the results to 10.

4.1.1.8 Map interface

The UI includes a map interface that uses OpenStreetMap. On the map the user can see the bounding box (if provided) with a yellow outline. The resulting datasets based on the user's query are expressed by blue boxes. These only occur if the metadata of these sources include longitude and latitude coordinates.



Figure 12 - Map interface using OpenStreetMap. The yellow box indicates the bounding box chosen by the user. The blue boxes are the resulting data sources found.

4.1.2 SPARQL API

Instead of using the search UI, users can also use the SPARQL API, which is available at <https://fair-ease-iddas.maris.nl/sparql> and makes use of Yagui (<https://docs.triply.cc/yasgui-api/>). The direct link to the SPARQL service can also be used, which is available at <https://fair-ease-iddas.maris.nl/sparql/query>. The SPARQL API supports GeoSPARQL for geospatial search.

In the SPARQL API, users can construct the SPARQL query themselves, or copy the SPARQL from the search UI and adapt it to their needs. Figure 13 illustrates the input field for the SPARQL query. The user can click on the triangle to execute the query and retrieve the results. The results are given in a table (Figure 14) with information on the title, media type, link to RDF representation, description, bounding box and catalogue. The results can also be downloaded to CSV.

```

Query × +
⚙️
1 PREFIX dcat: <http://www.w3.org/ns/dcat#>
2 PREFIX dc: <http://purl.org/dc/terms/>
3 PREFIX prov: <http://www.w3.org/ns/prov#>
4 PREFIX xsd: <http://www.w3.org/2001/XMLSchema#>
5 PREFIX geof: <http://www.opengis.net/def/function/geosparql/>
6 PREFIX geo: <http://www.opengis.net/ont/geosparql#>
7 PREFIX schema: <https://schema.org/>
8
9 SELECT DISTINCT (GROUP_CONCAT(DISTINCT ?_title; SEPARATOR=' ') AS ?title) (GROUP_CONCAT(DISTINCT ?_mediaType; SEPARATOR='|') AS ?mediaType) ?dataset ?
description ?bbox (REPLACE(STR(?catalog), '\\?.*$', '')) AS ?cleanCatalog
10 WHERE {
11
12   ?dataset a dcat:Dataset ;
13           dc:title ?_title ;
14           dc:description ?description .
15
16   OPTIONAL {
17     ?dataset dc:temporal [
18       a dc:PeriodOfTime ;
19       dcat:startDate ?startDate ;
20       dcat:endDate ?endDate
21     ] .
22   }
23
24   OPTIONAL {
25     ?dataset dc:spatial [
26       a dc:Location ;
27       dcat:bbox ?bbox
28     ] .
29   }
30
31   OPTIONAL {
32     ?dataset schema:variableMeasured [
33       a schema:PropertyValue ;
34       schema:name ?parameterName
35     ] .
36

```

Figure 13 - SPARQL API input field.

Table Response 10 results in 8.568 seconds

Simple view Ellipse Filter query results Page size: 50

title	mediaType	dataset	description	bbox	cleanCatalog
1 Black S...	https://www....	<http://dat...	EMODnet Chemistry aims to provide access to marine chemistry data sets and derived data produ...	"POLYGON ...	http://data.blu...
2 Copep...	https://www....	<http://dat...	Gegevens over densiteiten en soortensamenstelling van copepoden op een sublitoraal zandig stat...	"POLYGON ...	http://data.blu...
3 LifeWat...	https://www....	<http://dat...	In the framework of the Lifewatch marine observatory a number of fixed stations on the Belgian Par...	"POLYGON ...	http://data.blu...
4 58G22...	https://www....	<http://dat...	ICOS OTC SOOP Release	"POLYGON ...	http://data.blu...
5 HELCO...	https://www....	<http://dat...	This dataset includes the data used in in the HELCOM/OSPAR Ballast Water Exemptions Decision ...	"POLYGON ...	http://data.blu...
6 Sample...	https://www....	<http://dat...	Longitudinal sampling of the rainbow trout (<i>Oncorhynchus mykiss</i>) microbiome reveals effects of di...	"POINT (-76...	http://data.blu...
7 Faunal ...	https://www....	<http://dat...	Occurrence (presence/absence) of the crabs at Kanamai Beach, Bamburi, Mkomani and the Gazi ...	"POLYGON ...	http://data.blu...
8 Sample...	https://www....	<http://dat...	lcWGS data for <i>Salvelinus alpinus</i>		http://data.blu...
9 26RA2...	https://www....	<http://dat...	ICOS OTC SOOP Release	"POLYGON ...	http://data.blu...
10 Algal c...	https://www....	<http://dat...	Data on the location of found macroalgal and seagrass species and study of the algal community o...	"POLYGON ...	http://data.blu...

Figure 14 - Results from a SPARQL query.

4.2 Data access

In the current version of the IDDAS, data access is possible for assets that are exposed via the DAB. By using the search UI, a set of data sources can be found. After clicking on a data source, a specific page is opened where access links to for example an FTP, WMS, or a direct download are provided. As an example in Figure 15, you can see for a data set from the SDI catalogue there are two links available to obtain the data.

O3, European air quality data for 2015, (interpolated data)

<https://essi-lab.eu/dab/fair-ease/dataset/ead3f70c-ba7e-4c2c-8073-f79a52bf1aae>

The data set provides concentrations for the air pollutants Ozone (O3) at 1 km grid combining monitoring air quality data in a 'regression-interpolation-merging mapping' methodology and the observational values of the air quality monitoring stations used in the interpolation. It covers Europe for the year 2015. It provides estimates for human health related indicators of pollutants ozone (93.2 percentile of maximum daily 8-hour means, SOMO35) and vegetation related ozone indicators (AOT40 for vegetation and for forests) for the year 2015.

Data distribution:

- (https://sdi.eea.europa.eu/webdav/datastore/public/eea_r_3035_1_km_aq-interpolated-O3_p_2015_v01_r00/)
- [Direct download \(https://sdi.eea.europa.eu/data/ead3f70c-ba7e-4c2c-8073-f79a52bf1aae\)](https://sdi.eea.europa.eu/data/ead3f70c-ba7e-4c2c-8073-f79a52bf1aae)



Figure 15 - Data access links to retrieve a data set found via the search UI.

To provide standardised data access via the IDDAS, we will identify and catalogue the data access protocols supported by the data sources (section 3.2.3.2), and integrate these with the DCAT-FE. This integration ensures that each dataset within the DCAT-FE references the protocol(s) that are available. By incorporating URIs directly into the metadata, we establish a standardised approach for documenting and accessing protocols. This way we can have plugins connecting to certain protocols, like assets with protocol STAC catalogue resource URL. The plugin could search those catalogues and be given resource URLs, data schemas that it can provide to the user.

Next to this, we will implement the Blue-Cloud API's ordering and access mechanism to enable the access to assets exposed via the Blue-Cloud API. We will expose an endpoint in the Blue-Cloud API, which uses a list of URIs to order the datasets and returns an order identifier to the UDAL/client, after which the UDAL/client awaits the order and can download the assets.

In the EAL we also want to provide a set of notebooks that can be used to access data via certain protocols such as WMTS, FTP. When a user finds a data product in the IDDAS that they want to use in their analysis, they retrieve a link that can be directly accessed from within the related notebook available in the EAL. The data product that is selected is then downloaded into their workspace, ready to be used.

5. Integration of IDDAS

5.1 Integration with FAIR-EASE Earth Analytical Lab (EAL)

The IDDAS is currently available from a separate server, but will in November be available from the FE D4sciences VRE (<https://fair-ease.d4science.org/>) as part of the FE EAL, illustrated in Figure 16. The asset selector will, like the current version of the IDDAS, allow for discovery and access to the assets harvested from the data infrastructures. But in addition, it will also be possible to push the assets into the user's local workspace in their virtual lab to conduct their analysis, or use for input to their models and services.

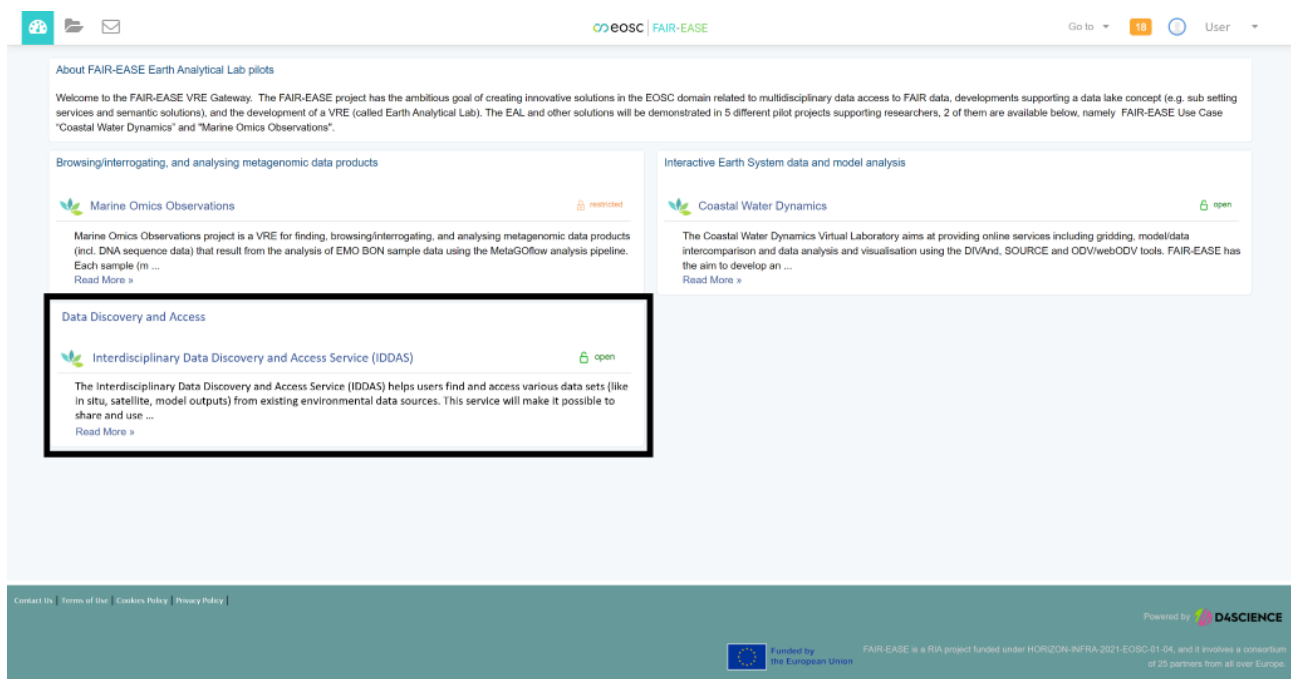


Figure 16 - Access point of IDDAS from the FAIR-EASE EAL.

The asset selector is foreseen to include a human readable as well as a machine accessible interface with similar functionalities as the IDDAS. For the design of the asset selector, we will involve pilots to make sure the selector serves the needs, and offers access to the required data. During the General Assembly (GA) planned in the beginning of October in Naples, we will have a working session around this topic.

A big challenge for efficient use of the IDDAS within the workflows of the pilots, is the availability of (meta)data as provided by the data infrastructures. The processing services in FE are often unable to work with the file collections as they are delivered by the data infrastructures. The files contain too many parameters, or the format is not fit as input. In reality, the services require as much as possible the correct subset of data, in a fit-for-purpose file format. Therefore, sub-setting services are extremely important to the end-user in the EAL, as they deliver exactly to the user what they need. For example, one single NETcdf file, containing all data they require to study a certain region for a specific time period and depth range. Instead of a large collection of separate files, for which they first need to conduct their own analysis in order to retrieve the relevant data.

Current solutions in our domains often don't provide these access services yet to the user. With the IDDAS we want to allow the user to access sub-setting services with predefined queries gaining easy access to the subsets they are looking for and pushing those into their workspace.

5.2 Connection with Uniform Data Access Layer (UDAL)

The UDAL aims to take advantage of the asset catalogue and the definitions of named queries in the data-demand registry. The idea is to make use of the information from data providers and the data demand, and negotiate the data and format with the IDDAS in order to execute the queries. The UDAL is a client-side abstraction layer with factories and plugins that can load (or negotiate) multiple (or alternative) implementations that all play the same access-dialogue-game. It is not a component of the IDDAS, but will make use of the IDDAS asset catalogue. One possible implementation/plugin inside FE should use the IDDAS for ultimate flexibility and should be easily replaceable with many others though (also outside FE). We will make choices on how to make use of the UDAL to enable the user or machine to access the datasets identified via the asset selector. We can then make several iterations in the IDDAS, adjusting the service and/or backend using the UDAL results. For more information on the UDAL, you can take a look at D4.3 - Status and expectations of the FAIR-EASE data lake.

The base domain used for the UDAL is <https://lab.fairease.eu/>, linking to a set of UDAL GitHub repositories:

- <https://github.com/fair-ease/py-udal-interface>: This includes the specification of the UDAL.
- <https://github.com/fair-ease/py-udal-fe-impl>: This includes an implementation example.
- <https://github.com/fair-ease/dataset-demand-register>: This includes the dataset demand register.

We are investigating multiple different approaches for linking the UDAL with the IDDAS. One example is, the application of UDAL with an asset from EMODnet Chemistry found in the IDDAS with the following link:

- <https://fair-ease-iddas.maris.nl/dataset?uri=https://data.blue-cloud.org/search/dcat/emodnet-chemistry/record?id=d352642b-e188-43d5-b7f7-df4a103badb0>

The IDDAS SPARQL API will be used to connect to the Blue-Cloud API and download the dataset. For this, we need to upgrade the Blue-Cloud API for it to work and need to:

1. Extend the dcat:Distribution info in the IDDAS SPARQL
2. Use this information to search for the chemistry record.
3. Go to the API to 'order' this record. (Note: This won't be an instant download, the Blue-Cloud DDAS is made to retrieve multiple datasets, add them to a .zip archive, and then return the data later. This process can take 10 mins to an hour.)
4. After the order is complete, UDAL should automatically serve this information to the user, without the user having the know about the steps mentioned here.

6. Sustainability of results

If we look towards the sustainability of the results, we can identify three related Key Exploitable Results (KERs), which each can be exploited separately in the EOSC framework:

- The FAIR-EASE DCAT profile (DCAT-FE)
- The newly developed software application, the “Semantic Analyser” (SA)
- The IDDAS, by providing implementation prototypes

All three results have been developed from scratch (TRL0). The plan for all results is to reach TRL7 at the end of the project. The results will be exploited via task-forces and collaboration initiatives, by showcasing the benefits of describing data assets with the DCAT-FE profile and using the SA for semantic harmonisation.

DCAT-FE (in principle generally applicable) will provide the option to enable machine to machine discovery of multi-disciplinary datasets, independent of datasets being published in a metadata catalogue or via subsetting services like ERDDAP, BEACON or other. By showing in FE via the IDDAS the associated benefits, we hope to promote wider adoption by incentivising data infrastructures to describe their (meta)data in DCAT-FE. This would then add to bridging the gap between multi-disciplinary data infrastructures in terms of data access. The profile is included on the FE GitHub, and will be maintained after the project ends. Using version numbering, the profile can be updated to include new (meta)data fields. Following a need for a cross-disciplinary (meta)data profile, DCAT-FE can possibly become a new community standard of describing datasets (assets) from multi-disciplinary data sources, used in e.g. EOSC. Bringing a solution that many EOSC developments can benefit from.

To deliver the level of semantic alignment described here, it requires a comprehensive understanding of the datasets targeted, their metadata elements and structure, and the semantic elements and concepts they contain. This is achieved with the help of a newly developed software application, the SA. The SA extracts textual information from selected metadata records and data files, and compares it to semantic artefacts (terms from ontologies and controlled vocabularies) held in a purpose-built FE Knowledge Base. Via BODC, it is planned to keep improving the SA in the context of other European projects. The integration of IDDAS with the semantic analyser enhances the FAIR principles, and WP6 intends to conduct a new evaluation to demonstrate the extent to which the FAIRness of these datasets has been enhanced through the use of IDDAS. This study is being conducted as part of development cycle 4, with the results to be included in deliverable D6.5: Guidelines for the improvement of the FAIRness of digital resources and services. Earlier, WP6 provided the results of the analysis of the FAIRness of selected datasets used by pilots (MS06 milestone). This analysis was conducted using the FUJI tool and the FMMM method.

The IDDAS will contain an asset selector that will allow users to search for relevant data using a faceted search based on keywords connected to the metadata and depending on the metadata-completeness. Using metadata, the user will have the flexibility to perform data selection on either a portion of a larger file, multiple files, a set of files, or any combination thereof. This choice is influenced by the user's specific needs and the processing service's capabilities. By testing and setting up this type of searching and accessing of multi-disciplinary data in FE, we can identify the improvements over traditional access catalogues and e.g. allow for uptake of these mechanisms via

Blue-Cloud as the EU marine operational component. This means that not the infrastructure itself, but rather the developments that showed improvements in terms of multi-disciplinary data access can be used and exploited in operational projects such as Blue-Cloud2026, EOSC etc.

7. Appendix

```

@prefix ex: <https://fairease.eu/catalogs> .
@prefix exdata: <https://fairease.eu/data> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix sosa: <http://www.w3.org/ns/sosa/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix dcat: <http://www.w3.org/ns/dcat#> .
@prefix dct: <http://purl.org/dc/terms/> .
@prefix geosparql: <http://www.opengis.net/ont/geosparql#> .
@prefix prov: <http://www.w3.org/ns/prov#> .
@prefix foaf: <http://xmlns.com/foaf/0.1/> .
@prefix dqv: <http://www.w3.org/ns/dqv#> .
@prefix adms: <http://www.w3.org/ns/adms#> .
@prefix sdo: <https://schema.org/> .
@prefix iop: <https://w3id.org/iadopt/ont/> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#>.

<http://vocab.nerc.ac.uk/collection/P01/current/TEMP681/> a skos:Concept, sosa:ObservedProperty,
iop:Variable;
skos:prefLabel "Temperature (IPTS-68) of the water body" .

# iop:hasObjectOfInterest <http://vocab.nerc.ac.uk/collection/S21/current/S21S027> ;
# iop:hasProperty <http://vocab.nerc.ac.uk/collection/S06/current/S0600160/> .

<http://vocab.nerc.ac.uk/collection/L27/current/ARGO_QC/> a skos:Concept;
skos:prefLabel "ARGO quality flags".

<https://edmo.seadatanet.org/report/43> a prov:Organization ;
rdfs:label "BODC".

<https://edmo.seadatanet.org/report/4614> a prov:Organization ;
rdfs:label "Argo".

<http://vocab.nerc.ac.uk/collection/P22/current/28/> a skos:Concept;
skos:prefLabel "Oceanographic geographical features" .

#The main catalogue
ex:FAIREASECatalog a dcat:Catalog ;
dct:title "FAIREASECatalog Catalog"@en ;
rdfs:label "FAIREASECatalog Catalog"@en ;
foaf:homepage <http://example.org/catalog> ;
dct:publisher <https://edmo.seadatanet.org/report/43> ;
dct:language <http://id.loc.gov/vocabulary/iso639-1/en> ;
#Contains other catalogues
dcat:catalog ex:SDNCatalog, ex:CopernicusCatalog .

#Describe each Catalogue with as much detail as FAIRE-EASE has and list their datasets. here I only
include one dataset for SDN ex:MyDataset

ex:SDNCatalog a dcat:Catalog ;
rdfs:label "SeaDataNet Catalogue";
dcat:dataset exdata:MyDataset.

ex:CopernicusCatalog a dcat:Catalog ;
rdfs:label "Copernicus Catalogue".

#Describing the main dataset
exdata:MyDataset a dcat:Dataset, sdo:Dataset;
dct:title "PROVOR-V JUMBO Profiling Float - 2903783 - Argo LOV";
dct:description "SeaDataNet is the Pan-European infrastructure for marine and ocean data management
and delivery services. ";

```

```

dct:identifier <https://cdi.seadatanet.org/report/681> ;
adms:identifier [
    rdf:type adms:Identifier;
    rdf:parseType "Resource";
    skos:notation "10.1000/182" ;
    adms:schemaAgency <https://registry.identifiers.org/registry/doi> ;
];
dcat:theme <http://vocab.nerc.ac.uk/collection/P22/current/28/> ;

dct:issued "2008-12-04";
dcat:distribution exdata:MyDataset-001-csv ;
dqv:hasQualityMeasurement exdata:measurement1;
dct:publisher <https://edmo.seadatanet.org/report/4614>;
dct:creator <https://edmo.seadatanet.org/report/43> ;

#Temporal Info
dct:temporal [
    a dct:PeriodOfTime ;
    dcat:startDate "1967-01-10"^^xsd:date ;
    dcat:endDate "2021-04-09"^^xsd:date ;
];

#Spatial Info
dct:spatial [
    a dct:Location ;
    dcat:bbox ""POLYGON((
3.053 47.975 , 7.24 47.975 ,
7.24 53.504 , 3.053 53.504 ,
3.053 47.975
))""^^geosparql:wktLiteral ;
];

sdo:variableMeasured [
    rdf:type sdo:PropertyValue;
    sdo:name "Temperature (IPTS-68) of the water body";
    sdo:alternateName "WC_temp68";
    sdo:propertyID <http://vocab.nerc.ac.uk/collection/P01/current/TEMPP681> ;
];

prov:wasGeneratedBy exdata:Activity1 .

exdata:Activity1 a prov:Activity;
prov:used <http://vocab.nerc.ac.uk/collection/L05/current/134/> ;
#Suggestion to separate the sensor from the platform as we currently do not hold this info
prov:used <http://vocab.nerc.ac.uk/collection/B76/current/B7600031/> .

#Separation of platform and sensor by commenting out the following:
<http://vocab.nerc.ac.uk/collection/L05/current/134/> a sosa:Sensor, prov:Entity, skos:Concept;
skos:prefLabel "water temperature sensor".
#sosa:observes <http://vocab.nerc.ac.uk/collection/P01/current/TEMPP681/> ;
#sosa:isHostedBy <http://vocab.nerc.ac.uk/collection/B76/current/B7600031/> .

<http://vocab.nerc.ac.uk/collection/B76/current/B7600031/> a sosa:Platform;
skos:prefLabel "National Oceanography Centre Autosub6000 autonomous underwater vehicle" .
#sosa:hosts <http://vocab.nerc.ac.uk/collection/L05/current/134/> .

##Distribution
exdata:MyDataset-001-csv a dcat:Distribution ;
dcat:downloadURL <http://dcat.example.org/files/001.csv> ;
dct:title "CSV distribution of imaginary dataset 001"@en ;
dct:title "distribuci3n en CSV del conjunto de datos imaginario 001"@es ;
dcat:mediaType <http://www.iana.org/assignments/media-types/text/csv> ;
dcat:accessService exdata:subset-service-001 ;
dcat:byteSize "5120"^^xsd:nonNegativeInteger .

```

```
#The quality info here
exdata:measurement1 a dqv:QualityMeasurement ;
dqv:isMeasurementOf <http://vocab.nerc.ac.uk/collection/L27/current/ARGO_QC/> ;
dqv:value "good"^^xsd:string .

#A service
exdata:subset-service-001
rdf:type dcat:DataService ;
dct:conformsTo <http://dcat.example.org/apidef/table/v2.2> ;
dct:type <https://inspire.ec.europa.eu/metadata-codelist/SpatialDataServiceType/invoke> ;
dcat:endpointDescription <http://dcat.example.org/api/table-005/capability> ;
dcat:endpointURL <http://dcat.example.org/api/table-001> ;
dcat:servesDataset exdata:MyDataset .
```

Figure A1 - DCAT-FE model.