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

D2.3 - FAIR-EASE semantic brokerage service

Work Package	**WP2 - Discovery, Access and FAIR Data services**
Lead Authors (Org)	**Alexandra Kokkinaki (NOC-BODC), Gwenaëlle Moncoiffé (NOC-BODC), Tjerk Krijger (MARIS), Enrico Boldrini (CNR), Peter Thijsse (MARIS)**
Contributing Author(s) (Org)	**Maria Luisa Chiusano (UNINA), Marc Portier (VLIZ)**
Due Date	**31.12.2023**
Date	**31.01.2024**
Version	**Final**

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)

Funded by the European Union

Versioning and contribution history

Version	Date	Author	Orcid ID	Notes
0.1	19.12.2024	Tjerk Krijger (MARIS) Peter Thijsse (MARIS)	0000-0002-1722-0523	Chapter 1 - IDDAS
0.2		Gwenaëlle Moncoiffé (NOC-BODC)	0000-0001-6559-4178	Chapter 2 - Semantic Analyser and analysis Appendix 5.2
0.3		Alexandra Kokkinaki (NOC-BODC)	0000-0001-8042-6391	Chapter 2 - Knowledge base Chapter 3 - Asset catalogue Appendix 5.1
0.4		Enrico Boldrini (CNR)		Chapter 4 - Producing DCAT target profile
0.5	25.01.2024	Maria Luisa Chiusano (UNINA) Marc Portier (VLIZ)		Review
0.8	26.01.2024	Gwenaëlle Moncoiffé (NOC-BODC)		Added section on Semantic brokerage to Chapter 1
1.0	29.01.2024	Gwenaëlle Moncoiffé (NOC-BODC), Alexandra Kokkinaki (NOC-BODC)		Final Revision
Final	31.01.2024	Clémentine FERRE (Neovia)		Final edition for submission

Table of Contents

Funded by the European Union

List of Figures

List of Tables

Terminology

Terminology/Acronym	Description
API	Application Programming Language
CDI	Common Data Index
CF	Climate and Forecast
CHEBI	Chemical Entities of Biological Interest
CMS	Content Management System
CSV	Comma-Separated Values
CTD	Conductivity-Temperature-Depth sensors package
DAB	Discovery and Access Broker
DCAT	Data Catalog Vocabulary
DOI	Digital Object Identifier
EAL	Earth Analytics Lab
EDMO	EUROPEAN DIRECTORY OF MARINE ORGANISATIONS
EIONET	European Environment Information and Observation Network
EFO	Experimental Factor Ontology
EML	Ecological Metadata Language
EnvO	Environmental Ontology
EOSC	European Open Science Cloud
FAIR	Findable; Accessible; Interoperable; Reusable
FE	FAIR-EASE
GA	Grant Agreement to the project
GCMD	Global Change Master Directory
I-ADOPT	InteroperAble Descriptions Of Observable Property Terminology
ICES	International Council for the Exploration of the Sea
IDDAS	Interdisciplinary Data Discovery and Access Service
INSPIRE	INfrastructure for SPatial InfoRmation in Europe
ISO	International Organization for Standardization
JSON	JavaScript Object Notation
KB	Knowledge Base
LD	Linked Data

Funded by the European Union

Terminology/Acronym	Description
M2M	Machine to Machine
NCBI	National Center for Biotechnology Information
NetCDF	Network Common Data Form
NVS	NERC vocabulary Server
OZCAR	Observatoires de la Zone Critique: Application et Recherche
QUDT	Quantity, Unit, Dimension and Type ontologies
RDA	Research Data Alliance
RDF	Resource Description Framework
ROR	Research Organisation Registry
SA	Semantic Analyser
SDMX	Statistical Data and Metadata eXchange
SSN/SOSA	Semantic Sensor Network/Sensor, Observation, Sample, and Actuator Ontology
SPARQL	SPARQL Protocol and RDF Query Language
UDAL	Uniform Data Access Layer
UI	User Interface
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
WMO	World Meteorological Organisation
WP	Work Package
XML	Extensible Markup Language

Funded by the European Union

Executive Summary

Semantic brokerage is a key component of the FAIR-EASE (FE) Interdisciplinary Data Discovery and Access Service (IDDAS), because the data access services behind the IDDAS are diverse in syntax (technical formats) and semantics (content/used vocabularies). In order to provide a user of the FAIR-EASE Virtual Research Environment (or Earth Analytical Labs) optimised discovery and access to the datasets, both syntactic and semantic brokerage is needed. A solution that many EOSC developments benefit from. As part of the metadata brokerage service, the FE semantic brokerage provides translation services between the various terminologies used by the diverse data sources. It enables semantically enhanced discovery of datasets and facilitates semantic interoperability through alignment.

To deliver this level of semantic alignment 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 "Semantic Analyser" (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.

The second part of the work focuses on a harmonised metadata profile following the DCAT Linked Data standard. This target profile (in principe generally applicable) will provide the option to enable machine to machine discovery of datasets, independent of datasets being published in a metadata catalogue (harmonised syntactically by the DAB) or via subsetting services like ERDDAP, BEACON or other. FAIR-EASE has undertaken research on how this difference can be overcome, and provide later implementation prototypes via its IDDAS.

The resulting federation of semantically connected data resources, will facilitate the sharing of multi-disciplinary datasets with the FE Uniform Data Access Layer (UDAL), and allow for interaction with the Earth Analytic Lab (EAL).

This Deliverable, the third one under WP2 'Discovery, Access and FAIR Data services', follows D2.1 (https://zenodo.org/records/7920551) that described a first level analysis of 19 data infrastructures relevant to FAIR-EASE and D2.2 (https://zenodo.org/records/8337393) which described the data access mechanisms available at these data infrastructures.

Funded by the European Union

1. Introduction to the IDDAS and Semantic Brokerage

1.1. The IDDAS general concept

The goal of the Interdisciplinary Data Discovery and Access service (IDDAS) is to enable easy access to multidisciplinary and aggregated datasets (*in situ* datasets, satellite datasets, omics experiments and model outputs) from a range of environmental data infrastructures relevant to the FAIR-EASE scientific pilot use-cases. These will be discoverable and downloadable using an asset selector user interface (UI) from within the Earth Analytical Lab (EAL). Figure 1 shows the location of the IDDAS within the complete FAIR-EASE architecture, including the connection to the EAL via the Uniform Data Access Layer (UDAL) that will use the Assets Selector to request assets from the Assets Catalogue in order to discover relevant data collections and retrieve the corresponding assets.

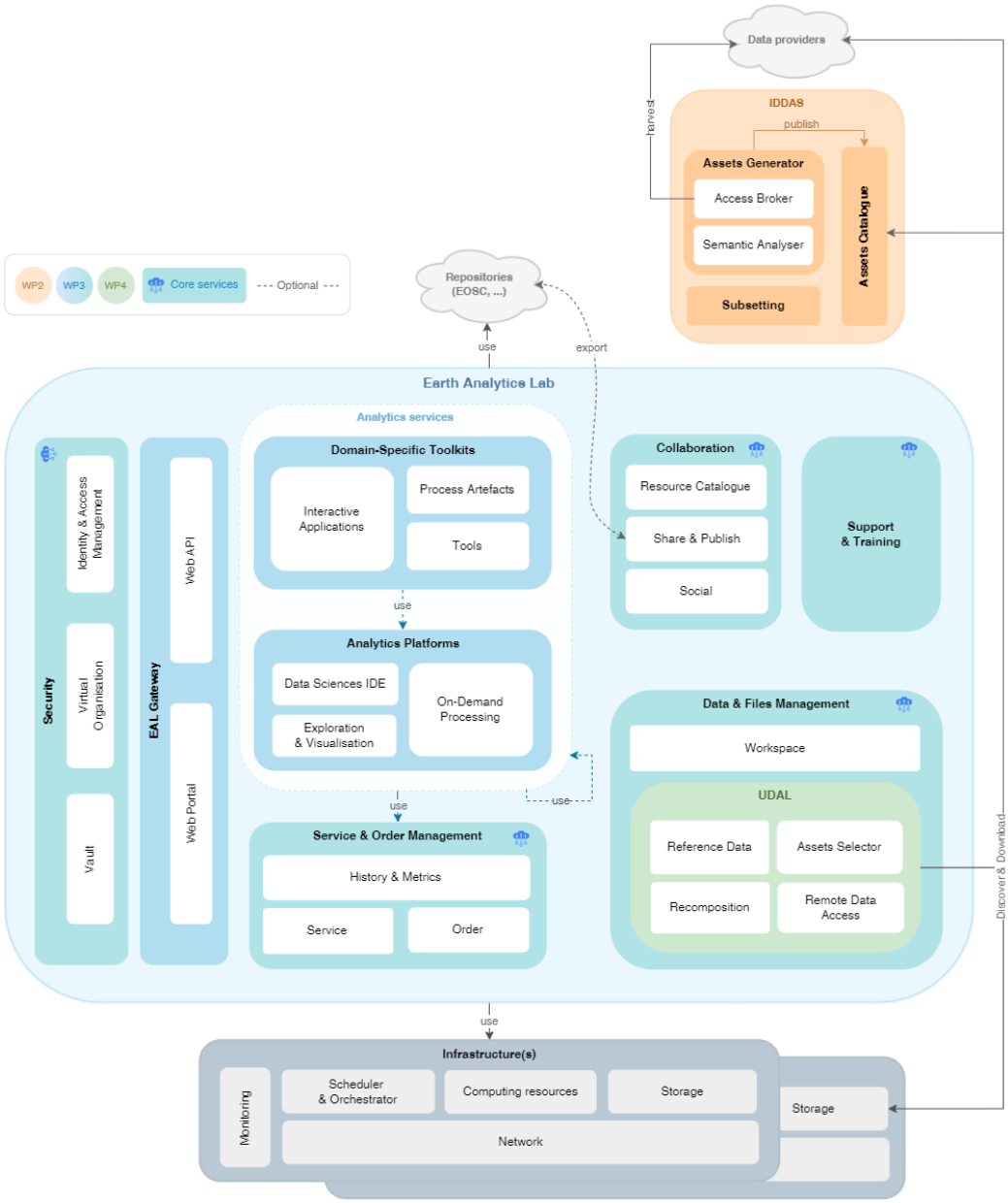

Figure 1 - Overall FAIR-EASE architecture

Funded by the European Union

For more information about the EAL see D3.1 (https://zenodo.org/records/10069773). The Assets Selector UI 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.

Behind the IDDAS there will be a catalogue of assets from which the user will be able to choose. This catalogue constitutes the semantic layer on top of three data access components:

- the FAIR-EASE DAB brokering framework
- FAIR-EASE cached data
- the services catalogue

The need for these three access components was identified following the analysis of the data infrastructures detailed in D2.1 (https://zenodo.org/records/7920551) and D2.2 (https://zenodo.org/records/8337393).

The DAB brokering framework harvests metadata from compatible data sources and harmonises them according to the the ISO 19115 XML profile based on the following common metadata elements: identifier, title, keyword, bounding box, temporal extent, parameter, instrument, platform, organisation, date-stamp, revision date, and resource identifier. The metadata will then be automatically converted to the proposed RDF DCAT standard described in Section 3 of this report. The semantic analyser described in Chapter 2 of this report will assist in detecting terms and controlled vocabularies used within the data sources. In particular, it will assist in identifying terms listed as keywords and associate them with the relevant FAIR-EASE metadata fields (e.g. 'platform', 'instrument', or 'parameter'). The output of the semantic analyser will also be used to identify useful and valuable mappings between terminologies in use in different domains, using a common set of reference vocabularies. This is an essential step to achieve greater cross-domain analytical capabilities.

Cached data access concerns datasets from data access services - like an FTP service - where in many cases there will be a lack of metadata to describe the data files harvested from services. Collections of datasets widely used by the different pilots can be cached (included as 'reference data' in the EAL) to improve access time, guarantee availability and allow for fast processing in the virtual labs. As an addition, there will probably also be a need for sub-setting services on top of these datasets, to improve usability in certain use cases, e.g. oxygen measurements in the top 10 metres of the ocean for the North Sea between 2000-2010. In the case of applying sub-setting services on the cached data, its offerings could be added to the IDDAS in DCAT standard. Any additional metadata information required will be added manually.

The third access component provides data access via the Services Catalogue. It will be used for data infrastructures with sub-setting services where related asset metadata might be lacking. Here, we can encourage data providers in control of the sub-setting services to publish in our proposed DCAT standard or alternatively, add them manually to the IDDAS.

The diagram in Figure 2 below shows the Asset Catalogue as a semantic layer between the three data access components mentioned above (DAB brokering framework, Data access services catalogue, Cached data) and the EAL.

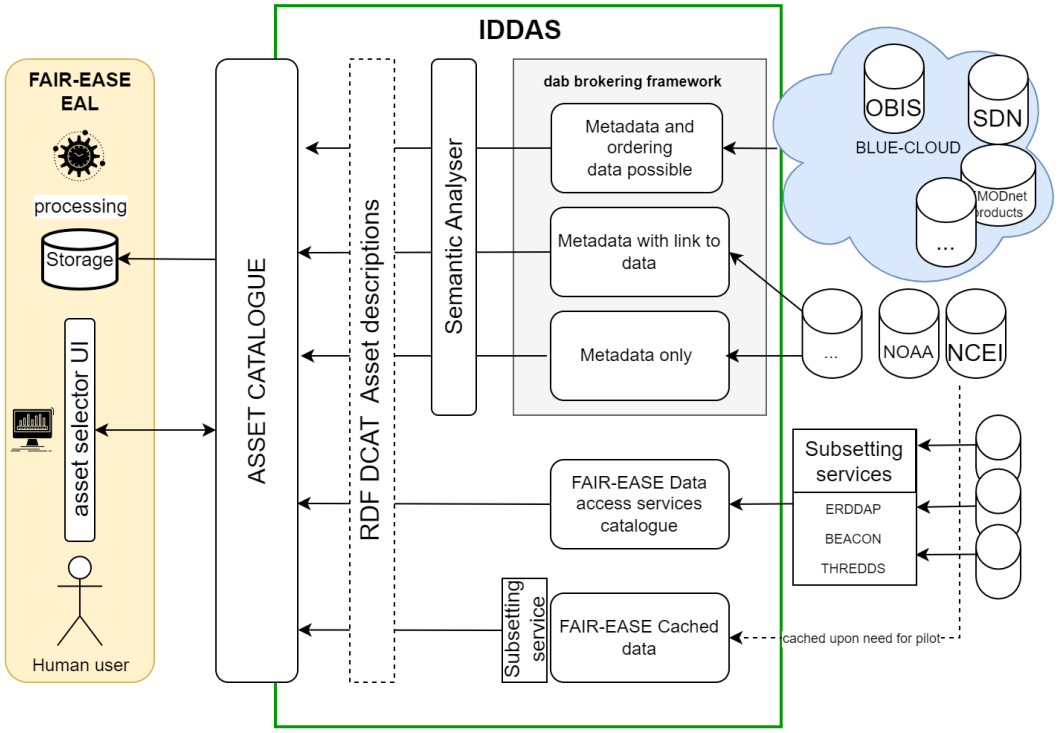

Figure 2 - Components of the IDDAS design

1.2. Semantic brokerage

A semantic brokerage service is necessary in order to complement syntactic brokerage and fully enable machine actionable tasks. For FAIR-EASE IDDAS and in particular its associated DAB, the goal is to semantically harmonise targeted metadata elements from incoming heterogeneous data sources and data files. The semantic broker needs to be able to perform three functions: 1) to unambiguously identify the targeted metadata concepts within a source file (metadata record or data file); 2) to identify connections or relationships between different terms; and 3) automatically provide translation services.

For example, within the earth environmental science domain, if a data source uses the term 'satellite' in its metadata records to refer to the observation platform from which the observation was made, and another data source uses the term 'Sentinel', the semantic broker will need to be able to 1) recognise that both are names for a type of observation platform used in environmental sciences, and 2) reconcile the fact that 'Sentinel' is a particular type of earth observation satellite; and 3) make the relationship available in a machine readable way.

Similarly if a data source describes a measurement as being 'chlorophyll' or 'chlorophyll-a' while another infrastructure describes a measurement as being 'phytoplankton pigments', the semantic broker will need to reconcile the fact that they might both be proxy measurements for plant biomass.

As a third example, a semantic broker should also be able to "know" that, if an instrument is identified as a 'CTD' (a common acronym for a Conductivity-Temperature-Depth sensors package ubiquitously used in oceanography) then, at a minimum, it will be associated with parameters related to the 'temperature' and 'salinity' of a water body. From this, one may also expect to be able to access the different types of units that may be used to express these properties of a water body.

These examples are depicting a reality that is not ideal since, in an ideal world, the content of data assets would be described using terms and URIs from known, well managed, and connected controlled vocabularies. However, while this may be achievable in the long-term and within specific scientific domains, there will always be a need for semantic brokerage to enable cross-domain interactions.

As mentioned above, for the FAIR-EASE IDDAS and DAB, the focus is on terms provided as keywords and terms used to describe three types of metadata elements: parameters, instruments, platforms. These are considered the minimum set of essential vocabulary-driven metadata elements that are common to most environmental observations from any scientific domains, and that can assist a user in selecting or excluding data based on high-level criteria. Information about these elements may be present as text with or without explicit reference to a particular controlled vocabulary, or/and as an identifier with or without unambiguous reference to a given controlled vocabulary.

To be able to provide semantic brokerage, a detailed semantic analysis of the datasets identified as relevant to the FAIR-EASE pilot use-cases, is necessary so that the necessary semantics used within all the data sources can be extracted. The semantic analysis process is described in the next chapter.

2. Semantic analysis

Extracting semantic artefacts (terms originating from ontologies, controlled vocabularies, thesauri) from the datasets and metadata records identified by the FAIR-EASE pilot use cases proved to be a laborious and time consuming task for the following reasons:

- Data spans a wide range of domains and infrastructures with semantic artefacts, when present, originating from numerous and diverse sources.
- Inconsistent and/or lack of referencing:
  - Terms are presented as free-text strings even though they may have come from controlled vocabularies published in dedicated semantic repositories
  - Some annotations contain typos or wrong labels or wrong URIs
  - Some annotations are located in the wrong metadata field
  - Some annotations are syntactically and semantically overly complex making it difficult to separate key semantic concepts or aligning them to known vocabularies.
- The information is missing.

In response to these challenges, we decided to create a new software application, the Semantic Analyser (SA), to help us in our analysis of FAIR-EASE semantic requirements and brokerage.

2.1. The Semantic Analyser (SA)

The SA is capable of extracting semantic content from metadata and data files encoded in various formats. In its first version, it supports netCDF data file formats and XML ISO 19115 harmonised metadata records coming from the FE DAB.

Upon completion of the analysis, the SA provides a list of terms that have been successfully extracted, along with information about the matched semantic artefacts.

2.1.1. Overall architecture

The SA has three main components

- A front-end UI (Figure 3)
- A codebase to read and extract semantic artefacts and compare them against a Knowledge Base (KB)
- The FAIR-EASE Knowledge Base

Funded by the European Union

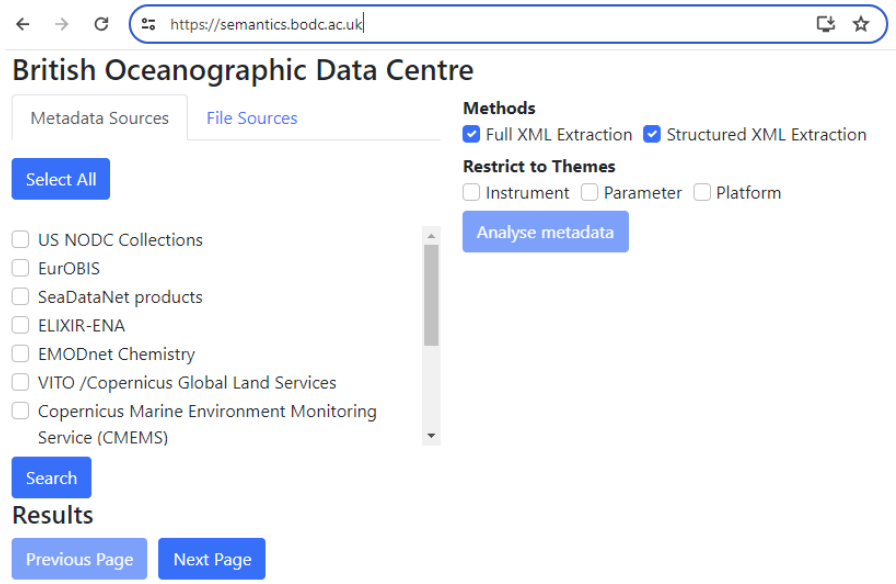

Figure 3 - Screenshot of the Semantic Analyser User Interface

Figure 4 shows the overall architecture and workflow of the SA, including uptake of XML and NetCDF file as input, checks against a knowledge base populated with terminologies from selected sources, and generation of results of the semantic analysis.

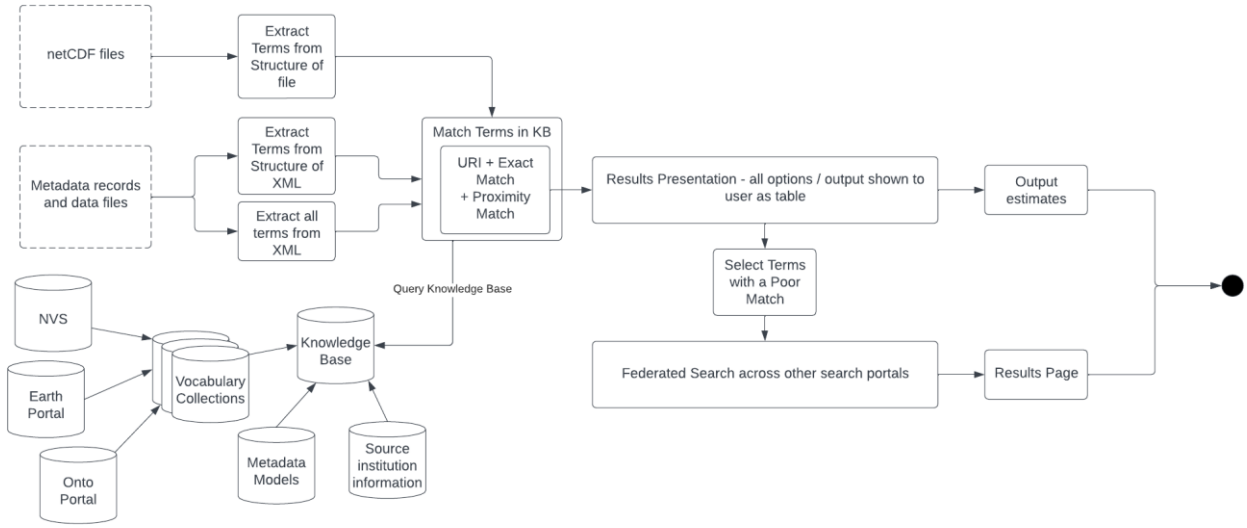

Figure 4 - Components and overall architecture of the Semantic Analyser

The next section describes how the SA functions in its current version 1.0.

2.1.2. SA Functionality

The frontend user interface (https://semantics.bodc.ac.uk/) enables users to select data or metadata assets to be analysed, switch on/off analytical methods, view and download results.

In this first version, the user interface implements the following design goals:

Funded by the European Union

- Users can select metadata records provided by the FAIR-EASE DAB
- One or more metadata records can be selected for concurrent analysis
- Users can view the raw XML of records selected for analysis
- The analysis can be restricted to different theme graphs in the triplestore
- The analysis method can be selected
- The results shall be displayed in a tabular format providing information on: the method and sub-method used; search and match terms; match URIs/links for the object with the property/object that matched; and the property that was matched on.
- the portion of search terms for which a result was found in the knowledge graph
- unmatched terms can be searched in 'federated' type manner in other analyser/search systems
- export of results in common data formats (CSV and JSON exports).

At this stage, input files can be either ISO 19115 XML metadata records or CF-netCDF. Only the metadata header of NetCDF files are read by the SA.

The content of the file is analysed with a current choice of three methods:

- Structured XML Analysis - based on the structure of ISO 19115 XML
- "Full" XML Analysis - which strips XML files of all textual fields
- NetCDF Analysis - which considers only the NetCDF metadata

2.1.3. The Knowledge Base

The knowledge base is composed of a number of ontologies and vocabularies which were identified as being used by or useful to FAIR-EASE partners and pilot cases. These form a set of reference data against which metadata or data elements can be matched. The ontologies and vocabularies have been organised into named graphs, with the named graph URIs typically being the namespace of the vocabulary or ontology, which allows easier organisation, searching, and updating. The ontologies and vocabularies have been categorised into the metadata elements (Keywords, Instruments, Parameters, and Platforms) that the terms within the vocabulary describe. The list of semantic resources is listed in the appendix in Chapter 5.

To enhance the matching results and expand the scope of the Knowledge Base, the Search Algorithm can also reach out to external terminology repositories in a federated search approach in order to find controlled vocabularies for terms that were present in the metadata file but not found in the FAIR-EASE Knowledge Base. This first version of the SA uses the APIs of two terminology repositories that could have vocabularies relevant to some of the FAIR-EASE pilots: BioPortal and EarthPortal. This functionality provides valuable insights into the semantic conventions employed by data providers and also helps identify additional terminology of interest. The newly discovered sources can then become a candidate for inclusion into the Knowledge Base.

Updating the Knowledge Base

Semantic artefacts are dynamic entities, evolving as new use cases emerge and as the underlying content and comprehension undergo changes. The frequency of updates varies, and regrettably, there is no consensus within the semantic community regarding the identification of the most recent version of a semantic resource for establishing an automated update process. In our approach, we split the update process into two distinct strands:

- Daily updates for controlled vocabularies from the NERC Vocabulary Server (NVS);
- The remaining Knowledge Base undergoes updates approximately twice a year, though the optimal frequency is still a subject of debate.

To facilitate daily updates for NVS, a compact Python application is employed to refresh the NVS vocabularies within the Fuseki knowledge base. This code utilises a SPARQL query to identify daily updates for any

vocabularies present in the Knowledge Base (KB). If new information is detected, the updater verifies its existence; if present, it is deleted, and if not, the new or updated triples are directly incorporated into the KB. A scheduled Continuous Integration (CI) job is executed each day to deploy and run this code.

To manage updates for the remaining knowledge base, we established a named graph, housing pertinent download URLs for each graph. The fuseki-updater, implemented as a Python application, retrieves the list of graphs and their respective download URLs from the database through a SPARQL query.

The data retrieved from these download URLs is stored in local files, which are then utilised to update the named graphs.

2.2. Preliminary results

This section provides a summary of our initial high level analysis of output from the Semantic Analyser run against XML files selected from data sources used by the FAIR-EASE pilot use-cases and available from the FAIR-EASE DAB at the time of the analysis.

General observations: data sources required by the pilot use-cases vary in their degree of semantic harmonisation. Some sources are harmonised across all the data products and datasets they provide, while others are not at all harmonised and may serve some datasets that use controlled vocabularies and some datasets that do not. There is also very little consistency in how these controlled vocabularies are used across the data assets they provide.

For the latter and for the FAIR-EASE project, it will be obviously important to focus our more detailed analysis on the datasets that matter to the FAIR-EASE pilots. This cannot currently be done directly using the SA UI as we cannot search resources within the SA interface as yet. However it is possible to analyse the XML file provided by the DAB using the SA XML upload functionality.

The table below (Table 1) summarises the results of our analysis of how the metadata related to Instruments, Platforms, Parameters and generic Keyword information is or could be made semantically accessible using a core set of controlled vocabularies. In this table, we list examples of terms found, the vocabularies to which we found some exact matches, suggest a set of recommended vocabularies to be used as reference for discovery and aggregation, and comment on additional possible steps to be taken to facilitate interoperability between the heterogeneous data sources. A more detailed summary table of our analysis of semantic annotation for individual data infrastructures is also included in the appendix in Chapter 5.

Table 1 - Analysis of results from the Semantic Analyser

Key to terminologies referenced: C17 (ICES Platform Codes), L05 (SeaDataNet device categories), L06 (SeaVoX Platform Categories), L22 (SeaVoX Device Catalogue), P01 (BODC Parameter Usage Vocabulary), P02 (SeaDataNet Parameter Discovery Vocabulary), P07 (Climate and Forecast Standard Names), P35 (EMODnet Chemistry aggregated parameter names), P36 (EMODnet Chemistry chemical groups), R25 (Argo sensor types), CHEBI (Chemical Entities of Biological Interest), EFO (Experimental Factor Ontology), EIONET (European Environment Information and Observation Network), EnvO (Environmental Ontology), GCMD (Global Change Master Directory Science Keywords), NCBI (National Center for Biotechnology Information),

Funded by the European Union

OZCAR (Observatoires de la zone critique, applications et recherche), QUDT (Quantity, Unit, Dimension and Type)

Metadata element	Examples of terms found in datasets	Vocabularies that return useful exact or related matches in the SA	Candidate vocabularies to be recommended as common reference	Additional recommendations
Instruments	Not often found as exact match although this is probably the easiest field to constrain against controlled vocabularies; also many occurrences of "not applicable" "unknown" or other unhelpful hits (e.g. deprecated platform code from C17)	NVS (multiple), GCMD (instruments), EFO	• L05/L22 could be used for marine and atmospheric domain • GCMD Instruments for other domains • EFO for biomolecular domain	Establish mappings between L05 and GCMD Instruments Investigate best options for instruments found in EFO
Platforms	Same as for instruments above; a few hits match "Satellite" "ships" but we also get hits for "not applicable" or "unknown"	NVS C17 and L06, GCMD	• L06/C17 could be used for marine domain • GCMD platforms	Establish mappings between L06 and GCMD platforms
Variables /Parameters	Some good examples of URIs usage to reference CF standard names (CMEMS); some exact matches to NCBITAXON, some close proximity matches P01, P02, S06, PATO, QUDT, Some exact match to sea areas like e.g. "black sea" and also "not applicable"	CF standard names, P01, P09, R03, P02, P35, P36 S06, L04, ncbitaxon, AphiaID, OZCAR, AERIS, GCMD,	• P07, P01 • R03, P09 • P02, P36, P35, GCMD, OZCAR • I-ADOPT[1]-compliant components representing the property, the object of interest, and the matrix	3 levels of information to be allowed: **1) Usage level parameter vocabularies** that are I-ADOPT compliant like e.g. P07, P01 or any usage level vocabularies that are mapped to I-ADOPT-compliant vocabs; 2) **Discovery level vocabularies** that are aligned or can be aligned to I-ADOPT-compliant usage vocabularies (P02, P36, others tbc) And 3) **I-ADOPT-compliant components of an observation**

[1] I-ADOPT is an ontology created by the Research Data Alliance (RDA) Working Group InteroperAble Descriptions of Observable Property Terminologies (Magagna et al, 2022); it enables the decomposition of complex variable names into a set of essential components that facilitates faceted discovery, interoperability and aggregation of datasets.

Funded by the European Union

				identifying either the property type (e.g. a QUDT quantity kind) or the object of the observation (e.g. a NCBI taxon or an AphiaID, a CHEBI ID, etc.)
Keywords	"Oceans", "Oceanographic geographical features" "Contaminants" "Orthoimagery" "biota" "environment" "soil moisture"	P05, P08, P22, GCMD, AERIS (often/always? duplicates GCMD), OZCAR, EIONET, EnvO	P05 (ISO19115 Topic Categories), P22 (Inspire themes), GCMD Science Keywords, Eionet, P08 (SeaDataNet Parameter Disciplines), Essential Variables and other relevant classes from EnvO.	The Keyword section concerns information that is not related to Instruments, Platforms or Parameters. Uses "top level" vocabularies. Mainly needed for discovery purpose and not so much for interoperability

As part of this exercise we identified a number of issues in the SA that will require fixing:

- Exact matches to CF standard names preferred labels are not recognised because underscores are currently removed prior to search; this issue is being fixed;
- Duplication of concepts between AERIS and GCMD vocabs due to AERIS integrating GCMD - Solution is to keep the GCMD version and only keep terms in AERIS that are not from GCMD;
- Need to highlight deprecated terms in mappings: e.g. terms that were deprecated in C17 because they were instruments or platform types and not platform instances should be highlighted in the UI and in the output;
- The SA seems to behave inconsistently when underscores are present resulting in missing exact matches; this issue is under investigation.

3. The FAIR-EASE Asset Catalogue

The IDDAS will require a comprehensive catalogue of assets (Figure 5) to manage assets like datasets, services, standards, software. To ensure that the assets adhere to FAIR principles, we are employing Linked Data and established vocabularies such as DCAT 3, SSN/SOSA, schema.org, Dublin Core, among others to describe them consistently. In this section, we are presenting the FAIR-EASE DCAT model (DCAT-FE) developed during the development cycle Dev Cycle 2. This will be used by the DAB broker to semantically enhance ISO 19115 records, and convert them to DCAT-FE, using output from the semantic analyser described in Chapter 2.

The DCAT-FE allows:

- The description of heterogeneous datasets using a standardised description that allows semantic annotation, while keeping the source ones intact;
- IDDAS to access a semantically and syntactically harmonised asset catalogue;
- The UDAL to aggregate such descriptions into a single point of access;

- The pilot researchers can easily find datasets from a single point of access.

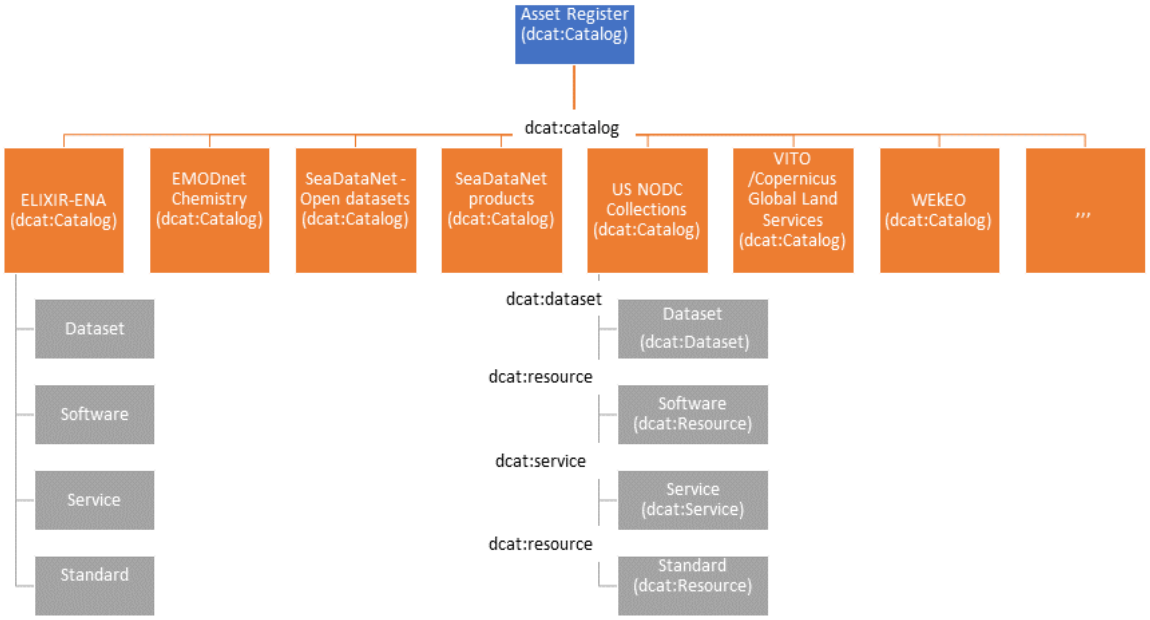

Figure 5 - The FAIR-EASE Asset Catalogue

3.1 Methodology

The data modelling took place during Dev Cycle 2 and was the result of an active collaboration with FAIR-EASE partners. The objective of the data modelling was to consistently describe all available datasets, aligning with the application requirements of the FAIR-EASE project. The uniform dataset description serves two main purposes:

- Semantic Searches: IDDAS utilises this information to facilitate semantic searches of the underlying data sources.
- Aggregation and Harmonisation: The described datasets play a crucial role in supporting the aggregation and harmonisation of data stored in the project's UDAL.

Our approach involved identifying metadata fields necessary for both describing and searching datasets.

3.2 The DCAT-FE mapping

The results of Dev Cycle 2 for dataset harmonisation are depicted in Table 2. Each metadata field is associated with a 'Description / Goal', 'Possible Value', 'DCAT property path' and 'DCAT turtle snippet'. The 'Possible values' column was derived from the semantic analysis of datasets performed using the SA that provided valuable insight into the semantic annotation of environmental datasets.

Table 2 - Results of Dev Cycle 2 Data Modelling

Funded by the European Union

Name of Field / Aspect	Description / Goal	Possible Value	DCAT property path	DCAT turtle snippet
Top level Keyword	**Description** Needs to be broader than just EV; need to be broad subject categories like e.g. gcmd keywords **GOAL** Enable flexible choice of discovery by broad "top level" description of contents of the dataset We can use a concept from a concept scheme or collection or an owl:Class High level parameters, genre	**P05 (ISO19115 Topic Categories), P22 (Inspire themes), GCMD Science Keywords, Eionet, P08 (SeaDataNet Parameter Disciplines), EnvO classes (including EVs)**	dcat:theme	dcat:theme <https://vocab.nerc.ac.uk/collection/P22/current/28/> ;
Parameters	**Description** Measured variables **GOAL** Discovery of actual narrow and well defined parameter semantics.	**3 types:** **Usage:** P01, P07, R03, P09, OZCAR Theia, EnvThes **Discovery/aggregated:** P02, P36, P35, A05 **Parameter description components:** NCBItaxon,WoRMS, ChEBI, or terms from GCMD, EIONET, or EnvO	**Option1:** prov:wasGeneratedBy/prov:used/sosa:observes **Option2:** sdo:variableMeasured	ex:Dataset a dcat:Dataset; prov:wasGeneratedBy dap:XXXX . dap:XXXX a prov:Activity; prov:used dap:SamplerA . dap:SamplerA a sosa:Sampler, prov:Entity; madeSampling ex:sample_01 . sosa:hasFeatureOfInterest {xxx} . dap:XXXX a prov:Activity; prov:used dap:Sensor1 . dap:Sensor1 a sosa:Sensor, prov:Entity; rdfs:type {instrument type} ; sosa:observes [a sosa: ObservedProperty, iadopt:Variable; iop:hasApplicableMatrix ex:Matrix; iop:hasApplicableProperty ex:Property; iop:hasApplicableObjectOfInterest ex:OOI .] sosa:isHostedBy {Platform} ; sosa:hasFeatureOfInterest {xxx} .

Funded by the European Union

Name of Field / Aspect	Description / Goal	Possible Value	DCAT property path	DCAT turtle snippet
				ex:Parameter
Date of publication	**Description** Date of formal issuance (e.g., publication) of the resource. **GOAL** Search with publication date		**dcterms:issued**	dcterms:issued "2011-12-05"^^xsd:date ;
Time period	**Description** Start-end of time spanned in the dataset **GOAL** Discovery of datasets/subsets that match the period-of-interest of the analysis		**dcterms:temporal/ dcat:startDate** **dcterms:temporal/ dcat:endDate**	dcterms:temporal [a dcterms:PeriodOfTime ; dcat:startDate "2016-03-04"^^xsd:date ; dcat:endDate "2018-08-05"^^xsd:date ;]. **For datasets with no end date** ex:ds127 a dcat:Dataset ; dct:temporal [a dct:PeriodOfTime ; dcat:startDate "2016-03-04"^^xsd:date ;] ; .
Spatial Description of the region of interest as a bounding box	**Description** Spatial Description of the region of interest as a bounding box **GOAL** Search using a bounding box		**dcterms:spatial/dcat:bbox**	dcterms:spatial [a dcterms: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 ;] .

Name of Field / Aspect	Description / Goal	Possible Value	DCAT property path	DCAT turtle snippet
Spatial Description of the region of interest as a polygon	**Description** A dataset whose spatial coverage is specified as a polygon (the coordinate reference system is CRS84).		**dcterms:spatial/locn:geometry**	dcterms:spatial [a dcterms:Location ; locn:geometry """POLYGON ((4.8842353 52.375108 , 4.884276 52.375153 , 4.8842567 52.375159 , 4.883981 52.375254 , 4.8838502 52.375109 , 4.883819 52.375075 , 4.8842353 52.375108))"""^^geosparql:wktLiteral ;] .
Spatial coverage of the dataset specified as a centroid	**Description** The geographic centre (centroid) of a spatial thing		**dcterms:spatial/dcat:centroid**	dcterms:spatial [a dcterms:Location ; dcat:centroid "POINT(4.88412 52.37509)"^^geosparql:wktLiteral ;] .
Data format & download URL	**Description** NetCDF, CSV, Zarr, Parquet, COG, ODV **GOAL** Discovery of the data format and link to the dataset	**https://www.iana.org/assignments/media-types/media-types.xhtml**	**dcat:distribution/dcat:mediaType** **dcat:distribution/downloadURL**	ex:dataset-001-csv dcat:distribution [a dcat:Distribution ; dcat:downloadURL <http://dcat.example.org/files/001.csv> ; dcterms:title "CSV distribution of imaginary dataset 001"@en ; dcat:mediaType <http://www.iana.org/assignments/media-types/text/csv> ;] .
Data collection/Data infrastructure	**Description** The Research Infrastructure (RI) this dataset is made available from **GOAL** Discovery datasets associated with this RI		**dcat:catalog**	ex:dataset dcat:catalog ex:ArgoCatalog . ex:ArgoCatalog a dcat:Catalogue .

Name of Field / Aspect	Description / Goal	Possible Value	DCAT property path	DCAT turtle snippet
Type of service	**Description** Returning complete data files, or subsets **GOAL** Discovery of the type of distribution of the dataset (e.g. "Give me the data services that are subsets", …)	**https://inspire.ec.europa.eu/metadata-codelist/SpatialDataServiceType**	**dcat:distribution/dcat:accessService/dcterms:type**	**ex:dataset-004** rdf:type dcat:Dataset ; dcat:distribution ex:dataset-004-csv . **ex:dataset-004-csv** rdf:type dcat:Distribution ; dcat:accessService ex:subset-service-005 ; dcat:accessURL <http://dcat.example.org/api/table-005> ; dcat:mediaType <http://www.iana.org/assignments/media-types/text/csv> . **ex:subset-service-005** rdf:type dcat:DataService ; dcterms:conformsTo <http://dcat.example.org/apidef/table/v2.2> ; dcterms: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-005> ; dcat:servesDataset ex:dataset-003.
unique identifier in the source	**Description** A unique identifier to refer to the dataset **GOAL** If the unique identifier is known, the user can choose the dataset using its unib que identifier	CDI and/or DOI,	dcterms:identifier and/or adms:identifier	ex:dataset dcterms:identifier "https://cdi.seadatanet.org/report/681"^^xsd:anyURI ; adms:identifier [a adms:Identifier; skos:notation "https://doi.org/10.5281/zenodo.1486279"^^xsd:anyURI; adms:schemaAgency <https://registry.identifiers.org/registry/doi> .] .
platform code	**Description** The code of the platform that the sensor was attached to when the measurement took place. **GOAL** *Discovery of*	WMO code, C17, ICES code	prov:wasGeneratedBy / prov:used / sosa:isHostedBy	ex:dataset a dcat:Dataset; prov:wasGeneratedBy dap:XXXX . dap:XXXX a prov:Activity; prov:used dap:Sensor1 . dap:Sensor1 a sosa:Sensor, prov:Entity; sosa:observes <P01> ; sosa:isHostedBy {platform code} .

Name of Field / Aspect	Description / Goal	Possible Value	DCAT property path	DCAT turtle snippet
	platforms associated with data collection			
instrument type	**Description** The type of instrument used for the measurement **GOAL** *Discovery of instruments used in data collection*	**L22, L05, gcmd instruments, ebi efo** **If L22 then L05 is inferred**	prov:wasGenerated By / prov:used / rdfs:type (or dcterms:type)	ex:dataset a dcat:Dataset; prov:wasGeneratedBy dap:XXXX . dap:XXXX a prov:Activity; prov:used dap:Sensor1 . dap:Sensor1 a sosa:Sensor, prov:Entity; rdfs:type {instrument type} ; sosa:observes <P01> ; sosa:isHostedBy ex:platform1 .
Quality flag	**Description** Flags used to provide additional information, usually referring to data quality, about data values **GOAL** Capture the quality flag scheme rather than the quality of the measurement (e.g. It has a 'good' quality according to that/this scheme)	L20 (quality terms), L27 (quality schemes) & others	dqv:hasQualityMea surement/ dqv:isMeasuremen tOf	ex:dataset dqv:hasQualityMeasurement :measurement1. :measurement1 a dqv:QualityMeasurement ; dqv:isMeasurementOf <dqv:isMeasurementOf>; dqv:value "https://vocab.nerc.ac.uk/collection/L20/current/4/" ^^xsd:anyURI .
Coordinate system	**Description** L10, … **GOAL** Discover the applicable coordinate reference system	http://www.opengis.net/def/crs/EPSG/0/28992 If not mentioned it is assumed that the coordinate reference system is CRS84.	dcterms:spatial/locn:geometry	ex:dataset dcterms:spatial [a dcterms:Location ; locn:geometry """**<http://www.opengis.net/def/crs/EPSG/0/28992>** POLYGON ((120749.725 487589.422 , 120752.55 487594.375 ,

Name of Field / Aspect	Description / Goal	Possible Value	DCAT property path	DCAT turtle snippet
		the geometry is always specified with WKT. As per [GeoSPARQL], when the CRS specification is omitted this implies that the default CRS is used - namely CRS84 (corresponding to WGS84, but with axis order longitude/latitude).		120747.735 487581.337 , 120751.564 487579.154 , 120755.411 487576.96 , 120750.935 487569.172 , 120755.941 487566.288 , 120764.369 487581.066 , 120749.725 487589.422))"""^^geosparql:wktLiteral ;] .
Institution ID (originator?)	**Description** The organisation/institution who **created** the resource **GOAL** Discovery of the institute "associated" to the dataset	EDMO, ROR	dcterms:creator	ex:dataset dcterms:creator <https://edmo.seadatanet.org/report/43> ;
Schema / Column information related to "Parameter" above	**Description** Describe the columns in the dataset **GOAL** Search for the column names of of the dataset		csvw:tableSchema	**ex:dataset-004** rdf:type dcat:Dataset ; dcat:distribution ex:dataset-004-csv . **ex:dataset-004-csv** rdf:type dcat:Distribution, csvw:Table ; dcat:accessService ex:subset-service-005 ; dcat:accessURL <http://dcat.example.org/api/table-005> ; csvw:tableSchema [Csvw:columns [csvw:name …. ; Dcterms:description … ; csvw:datatype …. ;] , […] ;];

Name of Field / Aspect	Description / Goal	Possible Value	DCAT property path	DCAT turtle snippet
Abstract/Description	**Description** A short description of the context / content of the dataset **GOAL** Discover datasets with lexical matching of free text		dcterms:abstract and/or dcterms:description	ex:dataset a dcat:Dataset; dcterms:abstract "{abstract}"^^xsd:string ; dcterms:description """{description}"""^^xsd:string ; .
Asset type	**Description** Provide the type of the asset **Goal** Discover certain types of assets e.g. datasets, services,	**dcat:Dataset**	**rdf:type**	

3.2 Future steps

This DCAT profile allows to describe environmental datasets originating from diverse sources and delivered harmonised by the FE DAB component. It would be important for the profile to be able to describe:

- A subset of an already existing dataset
- A derived dataset

Additionally the profile deals only with one type of asset, the dataset. The extended profile will be able to describe all asset types supported by IDDAS. Finally to enable automated validation, we will investigate the option of offering a SHACL specification for the DCAT- FE.

4. Producing DCAT target profile

4.1. 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 FAIR-EASE to perform syntactic harmonisation of the heterogeneous metadata documents made available by the different sources and enabling on top of them uniform discovery capabilities to the benefit of FAIR-EASE users.

The diagram in Figure 6 depicts the deployment of the FE DAB component. The information flow is started by the data providers (three sample providers occur in the diagram, while the actual growing list of providers currently holds 13 providers). Data providers 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.). The FE DAB acts as an interoperability enabler, connecting FE data providers to FE data consumers, by harmonising the different metadata descriptions towards a central harmonised one, based on ISO 19115 and supporting custom community extensions. Finally metadata is again converted according to different user requirements enabling the information flow to reach the user tools and applications. In particular, currently three clients of the FE DAB are identified:

- SA: gathers metadata from FE DAB CSW AP ISO interface to further analyse their semantic content and perform semantic augmentation of metadata records. Metadata documents in this cases are encoded using ISO 19139 schema
- Asset catalogue: provides a list of all available FE resources. The resources made available by the FE DAB are in this case encoded according to the DCAT-FE metadata profile.
- Test portal: a test portal meant to support simple user queries to assess the current content of the FE DAB. The test portal uses JSON as its data model for data discovery and result evaluation.

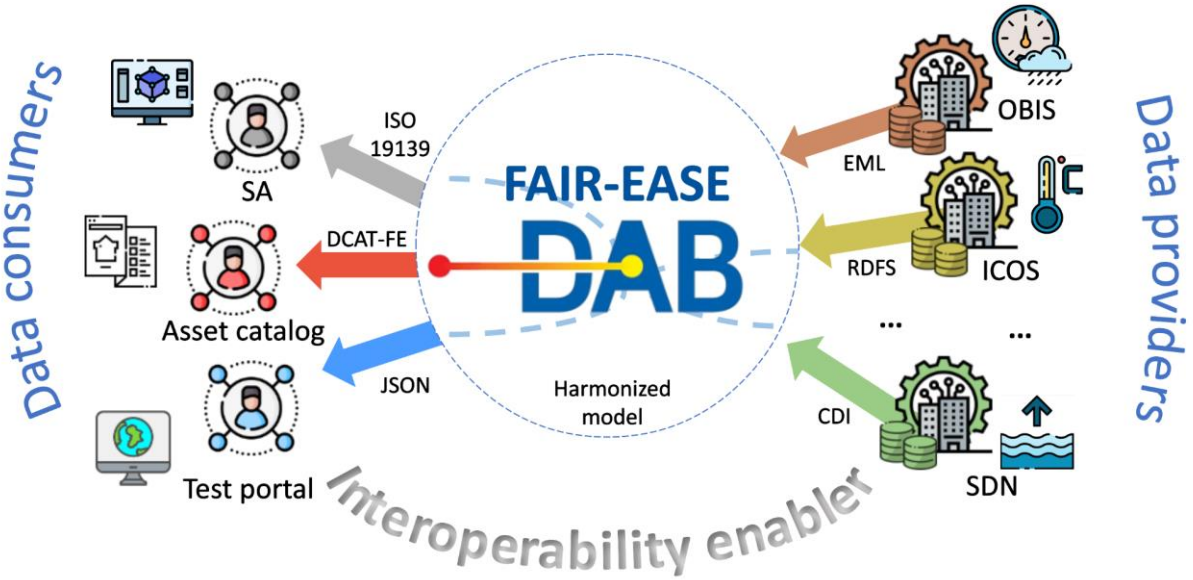

Figure 6 - The FE DAB common model

Each source makes available metadata descriptions according to different standards. FE DAB makes use of a common model to syntactically harmonise the available information and make it available according to different metadata encodings for further exploitation by the FE clients.

4.2. Mapping to DCAT-FE metadata model

A mapping from the DAB harmonised model towards DCAT-FE is being implemented in order to populate the FE asset catalogue with the resources made available from the FE DAB.

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 7 shows the mapping procedure occurring in the FE DAB. A first metadata mapping takes place starting from the original metadata harvested by the FE broker from each FE provider. Each

Funded by the European Union

metadata element is syntactically translated to correspondent elements in the DAB harmonised metadata model. A second mapping takes place starting from the harmonised metadata model, and has the target metadata information model as the output (for example the DCAT-FE model, as required by the FE asset catalogue).

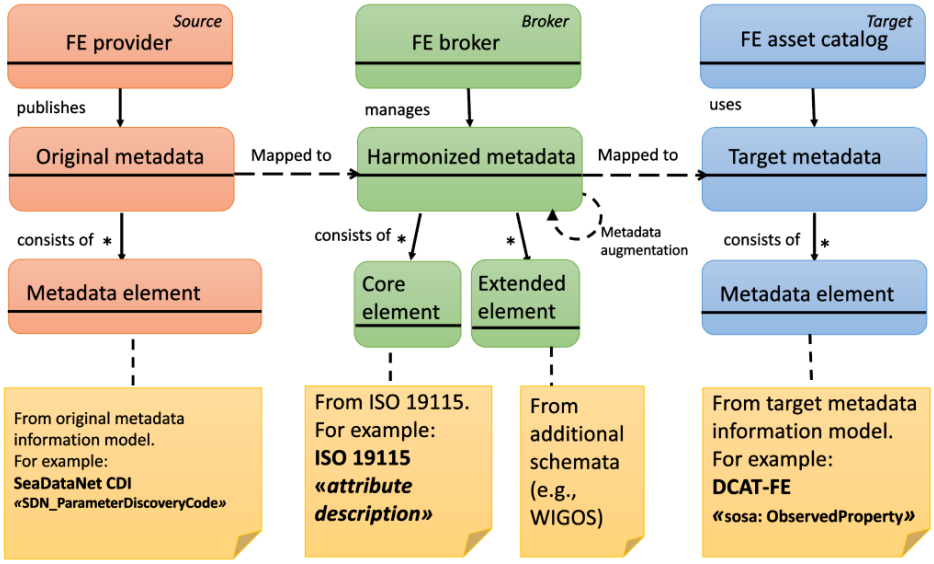

Figure 7 - Mappings occurring in the FE DAB

The following table (Table 3) provides detail of this second mapping from metadata elements of the FE DAB harmonised model based on ISO 19115 to the main DCAT-FE elements.

Table 3 - Metadata elements of FE DAB in relation to DCAT-FE elements

Name of Field / Aspect	**ISO 19139 property path**	**ISO 19139 snippet**
Top level Keyword	gmd:descriptiveKeywords	<gmd:descriptiveKeywords> <gmd:MD_Keywords> <gmd:keyword> <gmx:Anchor xlink:href="https://vocab.nerc.ac.uk/collection/P22/current/28/" xlink:title="Oceanographic geographical features">Oceanographic geographical features</gmx:Anchor> </gmd:keyword> <gmd:type> <gmd:MD_KeywordTypeCode codeSpace="SeaDataNet" codeList="..." codeListValue="theme" >theme</gmd:MD_KeywordTypeCode> </gmd:type> <gmd:thesaurusName> <gmd:CI_Citation> <gmd:title> <gco:CharacterString>GEMET - INSPIRE themes, version 1.0</gco:CharacterString> </gmd:title> <gmd:date>

Name of Field / Aspect	ISO 19139 property path	ISO 19139 snippet
		<gmd:CI_Date> <gmd:date> <gco:Date>2008-06-01</gco:Date> </gmd:date> <gmd:dateType> <gmd:CI_DateTypeCode codeList="..." codeListValue="publication" codeSpace="ISOTC211/19115" >publication</gmd:CI_DateTypeCode> </gmd:dateType> </gmd:CI_Date> </gmd:date> </gmd:CI_Citation> </gmd:thesaurusName> </gmd:MD_Keywords> </gmd:descriptiveKeywords>
Parameters	gmd:attributeDescription	<gmd:attributeDescription> <gco:RecordType xlink:href="SDN:P01::RFDSCH01" xlink:title = “Riverine discharge of water“>Riverine discharge of water</gco:RecordType> </gmd:attributeDescription>
Date of publication	Data identification gmd:date	<gmd:date> <gmd:CI_Date> <gmd:date> <gco:Date>2021-12-08</gco:Date> </gmd:date> <gmd:dateType> <gmd:CI_DateTypeCode codeList="..." codeListValue="revision" codeSpace="ISOTC211/19115" >revision</gmd:CI_DateTypeCode> </gmd:dateType> </gmd:CI_Date> </gmd:date>
Time period	gmd:temporalElement	<gmd:temporalElement> <gmd:EX_TemporalExtent> <gmd:extent> <gml:TimePeriod gml:id="mik3.7" > <gml:beginPosition>2009-01-01T00:00:00</gml:beginPosition> <gml:endPosition>2009-06-24T17:36:01</gml:endPosition> </gml:TimePeriod> </gmd:extent> </gmd:EX_TemporalExtent>

Name of Field / Aspect	ISO 19139 property path	ISO 19139 snippet
		</gmd:temporalElement> **For datasets with no end date:** <gmd:temporalElement> <gmd:EX_TemporalExtent> <gmd:extent> <gml:TimePeriod gml:id="mik3.7" > <gml:beginPosition>2009-01-01T00:00:00</gml:beginPosition> <gml:endPosition indeterminatePosition="now" /> </gml:TimePeriod> </gmd:extent> </gmd:EX_TemporalExtent> </gmd:temporalElement>
Spatial Description of the region of interest as a bounding box	gmd:geographicElement	<gmd:geographicElement> <gmd:EX_GeographicBoundingBox> <gmd:westBoundLongitude> <gco:Decimal>-68.548849</gco:Decimal> </gmd:westBoundLongitude> <gmd:eastBoundLongitude> <gco:Decimal>-49.007153</gco:Decimal> </gmd:eastBoundLongitude> <gmd:southBoundLatitude> <gco:Decimal>59.400296</gco:Decimal> </gmd:southBoundLatitude> <gmd:northBoundLatitude> <gco:Decimal>73.889864</gco:Decimal> </gmd:northBoundLatitude> </gmd:EX_GeographicBoundingBox> </gmd:geographicElement>
Spatial Description of the region of interest as a polygon	gmd:geographicElement	<gmd:geographicElement> <gmd:EX_BoundingPolygon> <gmd:polygon> <gml:MultiCurve gml:id="mc01" > <gml:curveMember> <gml:LineString gml:id="ls01" > <gml:description>This is line 1</gml:description> <gml:name>line1</gml:name> <gml:posList>-68.548849 73.889864 -61.408617 72.824456 -58.026401 68.136664 -56.523193 62.38344 -49.007153 59.400296</gml:posList> </gml:LineString> </gml:curveMember>

Name of Field / Aspect	ISO 19139 property path	ISO 19139 snippet
		<gml:curveMember> <gml:LineString gml:id="ls02" > <gml:description>This is line 2</gml:description> <gml:name>line2</gml:name> <gml:posList>112.963503 17.636232 114.842479 14.866168 114.842479 11.030688 111.084463 8.473704 106.574831 5.916728</gml:posList> </gml:LineString> </gml:curveMember> <gml:curveMember> <gml:LineString gml:id="ls03" > <gml:description>This is line 3</gml:description> <gml:name>line3</gml:name> <gml:posList>-76.816333 -36.777203 -77.192141 -43.722611 -80.198605 -47.195299 -78.319565 -50.400867 -74.561741 -54.674947 -70.052045 -57.079131</gml:posList> </gml:LineString> </gml:curveMember> </gml:MultiCurve> </gml:polygon> </gml:EX_BoundingPolygon> </gml:geographicElement>
Spatial coverage of the dataset specified as a centroid	Not available in ISO 19115 base line	
Data format	gmd:distributionFormat	<gmd:distributionFormat> <gmd:MD_Format> <gmd:name> <sdn:SDN_FormatNameCode codeSpace="SeaDatanet" codeListValue="ODV" codeList="https://vocab.nerc.ac.uk/isoCodelists/sdnCodelists/cdicsrCodeList.xml#SDN_FormatNameCode" >Ocean Data View ASCII input</sdn:SDN_FormatNameCode> </gmd:name> <gmd:version> <gco:CharacterString>0.3</gco:CharacterString> </gmd:version> </gmd:MD_Format> </gmd:distributionFormat>

Name of Field / Aspect	ISO 19139 property path	ISO 19139 snippet
Data collection/Data infrastructure	Not available in ISO 19115 base line	
Depth/height range	gmd:verticalElement	<gmd:verticalElement> <gmd:EX_VerticalExtent> <gmd:minimumValue> <gco:Real>0</gco:Real> </gmd:minimumValue> <gmd:maximumValue> <gco:Real>100</gco:Real> </gmd:maximumValue> <gmd:verticalCRS>[...]</gmd:verticalCRS> </gmd:EX_VerticalExtent> </gmd:verticalElement>
Type of service	Not available in ISO 19115 base line	
Unique identifier in the source	Data identification citation gmd:identifier	<gmd:identifier> <gmd:MD_Identifier> <gmd:code> <gco:CharacterString>ERD_EP_RVDS_INSITU</gco:CharacterString> </gmd:code> </gmd:MD_Identifier> </gmd:identifier>
Platform code	gmd:descriptiveKeywords	<gmd:descriptiveKeywords> <gmd:MD_Keywords> <gmd:keyword> <gmx:Anchor xlink:href="https://vocab.nerc.ac.uk/collection/C17/current/06AQ/" xlink:title="Polarstern">Polarstern</gmx:Anchor>

Funded by the European Union

Name of Field / Aspect	ISO 19139 property path	ISO 19139 snippet
		</gmd:keyword> <gmd:type> <gmd:MD_KeywordTypeCode codeSpace="SeaDataNet" codeList="..." codeListValue="theme" >platform</gmd:MD_KeywordTypeCode> </gmd:type> <gmd:thesaurusName> <gmd:CI_Citation> <gmd:title> <gco:CharacterString>C17 (ICES Platform Codes)</gco:CharacterString> </gmd:title> </gmd:CI_Citation> </gmd:thesaurusName> </gmd:MD_Keywords> </gmd:descriptiveKeywords>
Instrument type	gmd:descriptiveKeywords	<gmd:descriptiveKeywords> <gmd:MD_Keywords> <gmd:keyword> <gmx:Anchor xlink:href="https://vocab.nerc.ac.uk/collection/L22/current/NETT0006/" xlink:title="Closing net - Barnes (1953)">Closing net - Barnes (1953)</gmx:Anchor> </gmd:keyword> <gmd:type> <gmd:MD_KeywordTypeCode codeSpace="SeaDataNet" codeList="..." codeListValue="theme" >platform</gmd:MD_KeywordTypeCode> </gmd:type> <gmd:thesaurusName> <gmd:CI_Citation> <gmd:title> <gco:CharacterString>L22 (SeaVox Device Catalogue)</gco:CharacterString> </gmd:title> </gmd:CI_Citation> </gmd:thesaurusName> </gmd:MD_Keywords> </gmd:descriptiveKeywords>
Quality flag	Not available in ISO 19115 base line	
Coordinate system	gmd:referenceSystemInfo	<gmd:referenceSystemInfo> <gmd:MD_ReferenceSystem> <gmd:referenceSystemIdentifier>

Name of Field / Aspect	ISO 19139 property path	ISO 19139 snippet
		<gmd:RS_Identifier> <gmd:authority>[…]</gmd:authority> <gmd:code> <gmx:Anchor xlink:href="http://vocab.nerc.ac.uk/collection/L10/current/4326/" xlink:title="World Geodetic System 84" >World Geodetic System 84</gmx:Anchor> </gmd:code> </gmd:RS_Identifier> </gmd:referenceSystemIdentifier> </gmd:MD_ReferenceSystem> </gmd:referenceSystemInfo>
Processing	Not available in ISO 19115 base line	
Institution ID (originator?)	gmd:citedResponsibleParty	<gmd:citedResponsibleParty> <gmd:CI_ResponsibleParty> <gmd:organisationName> <gmx:Anchor xlink:href="https://edmo.seadatanet.org/report/18" xlink:title="Scott Polar Research Institute (SPRI)" >Scott Polar Research Institute (SPRI)</sdn:SDN_EDMOCode> </gmd:organisationName> <gmd:contactInfo>[...]</gmd:contactInfo> <gmd:role> <gmd:CI_RoleCode codeList="..." codeListValue="originator" codeSpace="ISOTC211/19115" >originator</gmd:CI_RoleCode> </gmd:role> </gmd:CI_ResponsibleParty> </gmd:citedResponsibleParty>
Schema / Column information	Not available in ISO 19115 base line	

Name of Field / Aspect	ISO 19139 property path	ISO 19139 snippet
related to “Parameter” above		
Abstract/Description	gmd:abstract	<gmd:abstract> <gco:CharacterString>EMODnet Physics - Collection of river flow rate (RVFL) TimeSeries - MultiPointTimeSeriesObservation</gco:CharacterString> </gmd:abstract>
Asset type	gmd:hierarchyLevel	<gmd:hierarchyLevelName> <gco:CharacterString>series</gco:CharacterString> </gmd:hierarchyLevelName>

As the table shows, most of the DCAT-FE elements can be mapped to and 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

Funded by the European Union

+ subTopicCategory

+ taxon

Additional keyword type codes could be defined as well, according to the FE requirements.

4.3. Subsetting services - BEACON Example

Beacon is a high-performance data lake solution used to store and subset millions of NetCDF datasets and terabytes of data with extensive query possibilities and fast retrieval. In the context of FAIR-EASE, it can be used to obtain subsets from a number of included data infrastructures. In order to map the output of Beacon to the DCAT target profile, the underlying data will be split up into multiple subsets based on region, parameter, time period and depth range. In the case of for example Euro-Argo, a subset could be for Oxygen in the North Sea, for a time period of one year for the whole depth column, for which there will be a direct download link available. And in this way a whole set of DCAT RDF will be created to a "blob" of asset descriptions. These subsets can then be found as assets in the asset selector within the EAL.

If the user requires a more dynamic query approach, by creating its own request, the Beacon system itself can be used. For more information regarding the use of Beacon, one can access the documentation here: https://beacon.maris.nl/

5. Appendix

5.1 Knowledge base content

List of semantic artefacts in the SA knowledge base

Vocabulary URL	Vocabulary Title
https://gcmd.earthdata.nasa.gov/KeywordViewer/scheme/chronounits	Chronounits
https://gcmd.earthdata.nasa.gov/kms/concepts/concept_scheme/DataFormat	DataFormat
https://gcmd.earthdata.nasa.gov/kms/concepts/concept_scheme/horizontalresolutionrange	Resolution range
https://gcmd.earthdata.nasa.gov/kms/concepts/concept_scheme/instruments	Instruments
https://gcmd.earthdata.nasa.gov/kms/concepts/concept_scheme/locations	Locations
https://gcmd.earthdata.nasa.gov/kms/concepts/concept_scheme/MeasurementName	Measurement Name
https://gcmd.earthdata.nasa.gov/kms/concepts/concept_scheme/MimeType	Mime Type

Vocabulary URL	Vocabulary Title
https://gcmd.earthdata.nasa.gov/kms/concepts/concept_scheme/platforms	Platforms
https://gcmd.earthdata.nasa.gov/kms/concepts/concept_scheme/projects	Projects
https://gcmd.earthdata.nasa.gov/kms/concepts/concept_scheme/providers	Providers
https://gcmd.earthdata.nasa.gov/kms/concepts/concept_scheme/rucontenttype	1112
https://gcmd.earthdata.nasa.gov/kms/concepts/concept_scheme/sciencekeywords	Science Keywords
https://gcmd.earthdata.nasa.gov/kms/concepts/concept_scheme/temporalresolutionrange	Temporal Resolution
https://gcmd.earthdata.nasa.gov/kms/concepts/concept_scheme/verticalresolutionrange	Vertical Resolution
http://inspire.ec.europa.eu/metadata-codelist/SpatialScope	Spatial scope
http://inspire.ec.europa.eu/metadata-codelist/SpatialScope/european	European
http://inspire.ec.europa.eu/metadata-codelist/SpatialScope/global	Global
http://inspire.ec.europa.eu/metadata-codelist/SpatialScope/local	Local
http://inspire.ec.europa.eu/metadata-codelist/SpatialScope/national	National
http://inspire.ec.europa.eu/metadata-codelist/SpatialScope/regional	Regional
http://purl.obolibrary.org/obo/envo.owl	The Environment Ontology
http://vocab.nerc.ac.uk/collection/C17/current/	ICES Platform Codes
http://vocab.nerc.ac.uk/collection/C30/current/	Active vocabulary content governance authorities
http://vocab.nerc.ac.uk/collection/E02/current/	Processing Levels for Earth Observing System Standard Data Products
http://vocab.nerc.ac.uk/collection/L05/current/	SeaDataNet device categories

Vocabulary URL	Vocabulary Title
http://vocab.nerc.ac.uk/collection/L06/current/	SeaVoX Platform Categories
http://vocab.nerc.ac.uk/collection/L22/current/	SeaVoX Device Catalogue
http://vocab.nerc.ac.uk/collection/P01/current/	BODC Parameter Usage Vocabulary
http://vocab.nerc.ac.uk/collection/P02/current/	SeaDataNet Parameter Discovery Vocabulary
http://vocab.nerc.ac.uk/collection/P03/current/	SeaDataNet Agreed Parameter Groups
http://vocab.nerc.ac.uk/collection/P05/current/	International Standards Organisation ISO19115 Topic Categories
http://vocab.nerc.ac.uk/collection/P06/current/	BODC-approved data storage units
http://vocab.nerc.ac.uk/collection/P07/current/	Climate and Forecast Standard Names
http://vocab.nerc.ac.uk/collection/P08/current/	SeaDataNet Parameter Disciplines
http://vocab.nerc.ac.uk/collection/P22/current/	GEMET - INSPIRE themes, version 1.0
http://vocab.nerc.ac.uk/collection/P24/current/	Units of measure dimensions
http://vocab.nerc.ac.uk/collection/P35/current/	EMODnet Chemistry aggregated parameter names
http://vocab.nerc.ac.uk/collection/P36/current/	EMODnet Chemistry chemical groups
http://vocab.nerc.ac.uk/collection/R01/current/	Argo data type
http://vocab.nerc.ac.uk/collection/R03/current/	Argo parameter codes
http://vocab.nerc.ac.uk/collection/R04/current/	Argo data centres and institutions
http://vocab.nerc.ac.uk/collection/R05/current/	Argo position accuracy
http://vocab.nerc.ac.uk/collection/R06/current/	Argo data state indicators

Funded by the European Union

Vocabulary URL	Vocabulary Title
http://vocab.nerc.ac.uk/collection/R07/current/	Argo history action codes
http://vocab.nerc.ac.uk/collection/R08/current/	Argo instrument types
http://vocab.nerc.ac.uk/collection/R09/current/	Argo positioning system
http://vocab.nerc.ac.uk/collection/R10/current/	Argo transmission systems
http://vocab.nerc.ac.uk/collection/R11/current/	Argo real-time quality-control test identifiers
http://vocab.nerc.ac.uk/collection/R12/current/	Argo history processing step codes
http://vocab.nerc.ac.uk/collection/R13/current/	Argo ocean area codes and boundary definitions
http://vocab.nerc.ac.uk/collection/R14/current/	Argo technical parameter names
http://vocab.nerc.ac.uk/collection/R15/current/	Argo trajectory measurement code identifiers
http://vocab.nerc.ac.uk/collection/R16/current/	Argo vertical sampling schemes
http://vocab.nerc.ac.uk/collection/R18/current/	Argo configuration parameter names
http://vocab.nerc.ac.uk/collection/R19/current/	Argo STATUS flags
http://vocab.nerc.ac.uk/collection/R20/current/	Argo GROUNDED flags
http://vocab.nerc.ac.uk/collection/R21/current/	Argo status flag on the Representative Park Pressure (RPP)
http://vocab.nerc.ac.uk/collection/R22/current/	Argo platform family
http://vocab.nerc.ac.uk/collection/R23/current/	Argo platform type
http://vocab.nerc.ac.uk/collection/R24/current/	Argo platform maker
http://vocab.nerc.ac.uk/collection/R25/current/	Argo sensor types

Funded by the European Union

Vocabulary URL	Vocabulary Title
http://vocab.nerc.ac.uk/collection/R26/current/	Argo sensor manufacturers
http://vocab.nerc.ac.uk/collection/R27/current/	Argo sensor models
http://vocab.nerc.ac.uk/collection/R28/current/	Argo controller board types and generations
http://vocab.nerc.ac.uk/collection/R40/current/	Argo Principal Investigator (PI) names
http://vocab.nerc.ac.uk/collection/RD2/current/	Argo delayed-mode quality control measurement flags
http://vocab.nerc.ac.uk/collection/RMC/current/	Argo measurement code categories
http://vocab.nerc.ac.uk/collection/RP2/current/	Argo profile quality control flags
http://vocab.nerc.ac.uk/collection/RR2/current/	Argo real-time quality control measurement flags
http://vocab.nerc.ac.uk/collection/RTV/current/	Argo float cycle timing variables
http://vocab.nerc.ac.uk/collection/S03/current/	BODC parameter semantic model sample preparation entity descriptions
http://vocab.nerc.ac.uk/collection/S04/current/	BODC parameter semantic model analytical method entity descriptions
http://vocab.nerc.ac.uk/collection/S05/current/	BODC parameter semantic model data processing entity descriptions
http://vocab.nerc.ac.uk/collection/S06/current/	BODC parameter semantic model parameter entity names
http://vocab.nerc.ac.uk/collection/S18/current/	BODC parameter semantic model physical entity names
http://vocab.nerc.ac.uk/collection/S19/current/	BODC parameter semantic model physical entity subgroup names
http://vocab.nerc.ac.uk/collection/S25/current/	BODC parameter semantic model biological entity names
http://vocab.nerc.ac.uk/collection/S27/current/	BODC parameter semantic model chemical substances

Funded by the European Union

Vocabulary URL	Vocabulary Title
http://www.linkedmodel.org/schema/vaem#GMD_QUDT-QUANTITY-KINDS-ALL	QUDT Quantity Kinds Version 2.1 Vocabulary
http://www.linkedmodel.org/schema/vaem#GMD_QUDT-UNITS-ALL	QUDT Units Version 2.1 Vocabulary
http://www.w3.org/1999/02/22-rdf-syntax-ns#	The RDF Concepts Vocabulary (RDF)
http://www.w3.org/2000/01/rdf-schema#	The RDF Schema vocabulary (RDFS)
http://www.w3.org/2002/07/owl	The OWL 2 Schema vocabulary (OWL 2)
http://www.w3.org/2004/02/skos/core	SKOS Vocabulary
http://www.w3.org/ns/sosa/	Sensor, Observation, Sample, and Actuator (SOSA) Ontology
http://www.w3.org/ns/ssn/	Semantic Sensor Network Ontology
https://vocab.aeris-data.fr/instrument	AERIS thesaurus: Instruments
https://vocab.aeris-data.fr/instrument	Thesaurus AERIS : Instruments
https://vocab.aeris-data.fr/parameter	Thesaurus AERIS : ParamÃ¨tres
https://vocab.aeris-data.fr/parameter	AERIS thesaurus: Parameters
https://vocab.aeris-data.fr/platform	Thesaurus AERIS : Plateformes
https://vocab.aeris-data.fr/platform	AERIS thesaurus: Platforms
https://vocab.aeris-data.fr/project	AERIS thesaurus: Projects
https://vocab.aeris-data.fr/project	Thesaurus AERIS : Projets
https://w3id.org/iadopt/ont	I-ADOPT Framework ontology
https://w3id.org/ozcar-theia	Theia/OZCAR thesaurus

Funded by the European Union

Vocabulary URL	Vocabulary Title
http://www.eionet.europa.eu/gemet	Labels and definitions in RDF
https://www.eionet.europa.eu/gemet/exports/latest/en/gemet-groups.rdf	Supergroups, groups and themes in RDF
<http://livercancer.imbi.uni-heidelberg.de/ccont>	cell culture ontology (CCONT)
<http://purl.bioontology.org/ontology/NCBITAXON/>	NCBI

Sparql query to find the list of graphs and their descriptions

```
PREFIX owl: <http://www.w3.org/2002/07/owl#>
PREFIX text: <http://jena.apache.org/text#>
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX dcat: <http://www.w3.org/ns/dcat#>

select ?a ?d ?g  where {

  {
     ?a <http://purl.org/dc/elements/1.1/title> ?d .

}
    union {
     { ?a <http://purl.org/dc/terms/title>    ?d .}
  }
   union {
     { ?a <https://gcmd.earthdata.nasa.gov/kms#viewer>    ?d .}
  }
   union {
     {
      ?a a owl:Ontology;
      <http://www.w3.org/2000/01/rdf-schema#label>    ?d .
       }
  }


}
```

**Some ontologies like efo although they contain a lot of metadata they do not include the ontology description:

```
select * where { graph <https://www.ebi.ac.uk/ols/ontologies/efo>
  {<http://www.ebi.ac.uk/efo/efo.owl> ?k ?l .}
} limit 1000
```

5.2 Summary of semantic analysis performed on a selection of metadata records in FAIR-EASE data sources

Compiled information derived from the semantic analysis of randomly selected data files from data sources used by the pilots in FAIR-EASE using the Semantic Analyser.

Key for Pilot cases: Coastal Waters Dynamics (5.1.1), Earth Critical Zone (5.1.2), Volcano Space Observatory (5.1.3), Ocean Biogeochemical Observations (5.2.1), Marine Omics Observations (5.3.1)

Key to terminologies referenced below: P07 (Climate and Forecast Standard Names), L06 (SeaVoX Platform Categories), L05 (SeaDataNet device categories), P02 (SeaDataNet Parameter Discovery Vocabulary), P35 (EMODnet Chemistry aggregated parameter names), P36 (EMODnet Chemistry chemical groups), R25 (Argo sensor types), R27 (Argo sensor models), C17 (ICES Platform Codes), GCMD (Global Change Master Directory Science Keywords), EnvO (Environmental Ontology), EIONET (European Environment Information and Observation Network), OZCAR (Observatoires de la zone critique, applications et recherche), NCBI (National Center for Biotechnology Information)

Data sources	Type and number of datasets (from D2.2 and D5.1 deliverables)	Result of high level semantic analysis of DAB harmonised XML	Pilots
Copernicus Marine Service (CMEMS)	There are currently 275 ocean data products available via the Marine Data Store that include hindcast, current and forecast data based on satellite observations, numerical models or in-situ observations. There are 16 ocean variables for which these products are available, ranging from nutrients to temperature to the carbonate system spanning the global ocean or local seas. **Datasets identified by pilots** ● 5.1.1: Original Temperature and Salinity observations; Ocean nutrients reanalysis; Surface Water Temperature; Surface Water Chlorophyll ● 5.2.1: Global Ocean Surface Carbon; Multi Observation Global Ocean 3D Temperature Salinity Height Geostrophic Current and MLD; Nutrient profiles vertical distribution; Global Ocean 3D Chlorophyll-a concentration, Particulate Backscattering coefficient and Particulate Organic Carbon; Global Ocean Colour Plankton and Reflectances MY L3 daily observations; Global Ocean Colour (CopernicusGlobColour), Bio-GeoChemical, L3 (daily) from Satellite Observations (1997-ongoing); Copernicus Marine In Situ - Global Ocean - Delayed Mode Biogeochemical Product'	**Parameter**: Good semantic description at usage level with what seems to be consistent use of CF Standard Names with NVS P07 URIs **Instrument**: no information **Platform**: no information **Generic keywords**: ISO 19115 Topics and INSPIRE themes e.g. "Oceans", "Oceanographic geographical features" **Comment**: ● Surprising occurrence of "not applicable" in the list of keywords; ● Dataset titles could be used as a source of additional keywords e.g. Global ocean; real time; in-situ observations; objective analysis; Arctic; ocean colour; plankton; transparency; L4; NRT; monthly observations; monthly mean; sea surface; wind and wind stress; scatterometer; model	**5.1.1** **5.2.1**
EMODnet Chemistry	EMODnet Chemistry provides access to marine chemical data, standardised harmonised validated	**Parameter**: Good semantic description at discovery/aggregation level with P02, P36, and	**5.1.1** **5.2.1**

Data sources	Type and number of datasets (from D2.2 and D5.1 deliverables)	Result of high level semantic analysis of DAB harmonised XML	Pilots
	data collections and reliable data products. EMODnet Chemistry focuses on four main themes: Eutrophication; Ocean acidification; Contaminants; Marine litter. Observational datasets include seawater quality, biota contamination, sediment quality and potential pollution. The measurement data cover 13 groups of chemical variables within all European sea regions. **Datasets identified by pilots** ● 5.1.1: Original nutrient observations; ● 5.2.1: Baltic Sea - Eutrophication and Acidity aggregated datasets 1902/2020 v2021; North East Atlantic Ocean - Eutrophication and Acidity aggregated datasets 1921/2020 v2021; Mediterranean Sea - Eutrophication and Acidity aggregated datasets 1911/2020 v2021; North Sea - Eutrophication and Acidity aggregated datasets 1921/2020 v2021; Arctic Ocean - Eutrophication and Acidity aggregated datasets 1923/2020 v2021; Black Sea - Eutrophication and Acidity aggregated datasets 1935/2020 v2021;	or P35 URIs and labels in the keyword section **Instrument**: no information **Platform**: no information **Generic keywords:** ISO 19115 Topics and INSPIRE themes **Comment**: ● information from titles could be mapped to CV e.g. North sea; dissolved oxygen; concentration; water body; mercury, sediment	
EMODnet Physics	● 5.1.1: Original, Practical Salinity (PSAL) Profiles from Mooring; Original, Sea Temperature (TEMP) Profiles from Mooring;	**NOT analysed as not yet available via the DAB**	**5.1.1**
EMODnet Biology	See EurOBIS	**See EurOBIS**	
Euro-Argo	Argo collects salinity/temperature profiles as well as biogeochemical profiles that contain Oxygen, Nitrate, Chlorophyll-a, pH measurements, suspended particles and downwelling irradiance, from an array of robotic floats that populate the ice-free seas and oceans that are deeper than about 2000 m, even 4000 for few Argo floats. Most profiles are made up of about 200 data points, but floats with high speed communications may be sending many more data points given the higher bandwidth. In total there are currently around 4000 Argo floats operational. **Datasets identified by pilots** ● 5.2.1: access to the entire dataset	**Parameter**: no exact matches, no URIs in the files we randomly selected but terms like "subsurface temperature", "subsurface salinity" and "subsurface pressure" are found. **Instrument**: Argo uses NVS R27 and R25 vocabularies for instruments using the **altLabel** for annotating metadata **Platform**: no match at present but could use a generic L06 or B76 entry or else R23 instances **Generic keywords:** No match found **Comment**: ● Argo has started using controlled vocabularies but these have not yet been implemented retrospectively to all the legacy data; ● The SA will help match terms to the	**5.2.1**

Data sources	Type and number of datasets (from D2.2 and D5.1 deliverables)	Result of high level semantic analysis of DAB harmonised XML	Pilots
		new vocabularies' URIs providing there are no typos in the text and an exact match can be made	
EurOBIS	Data on temporal and spatial distribution of marine species (angiosperms, benthos, birds, fish, macroalgae, mammals, phytoplankton, reptiles, zooplankton) and species traits from European regional seas as defined by the EEA's Europe's seas' dataset (Arctic Ocean, (North) Atlantic Ocean, Baltic Sea, Black Sea, Mediterranean Sea and North Sea). Measurements are focussed on biological occurrences: taxonomic identifications, the geotemporal information, the bundances /masses/counts. Other parameters provided include abiotic measurements (e.g. temperature, grain size), environmental facts (e.g. habitat), biotic measurements (e.g. abundance, length), biotic descriptors (e.g. lifestage, sex) and sampling descriptors (e.g. sampling instrument, surface area). Gridded products derived from the data in EMODnet are provided as well as the observation data itself. **Datasets identified by pilots** ● 5.3.1: Ecological time series data	**Parameter**: some exact and proximity matches however heterogeneous use of semantics within network **Instrument**: none found but could be due to heterogeneous use of semantics within network **Platform**: as above **Generic keywords**: none found **Comment**: ● Requires more targeted analysis	**5.3.1**
SeaDataNet CDI	The CDI service provides online access to a large number of marine and ocean data sets, managed by more than 117 connected SeaDataNet data centres originating from 34 countries around the European seas. Currently it gives access to more than 2.8 Million data sets, originating from more than 945 organisations in Europe, covering physical, geological, chemical, biological and geophysical data, and acquired in European waters and global oceans. **Datasets identified by pilots** ● 5.3.1: Physical oceanographic data measurements: instrument parameter measurements	**Parameter**: Good semantic description at discovery level with P02 **Instrument**: Good semantic description using L05 terms and identifiers **Platform**: Good semantic description using L06 terms and identifiers **Generic keywords**: ISO 19115 Topics and INSPIRE themes	**5.3.1**
VITO/Copernicus Land Monitoring Service	The Land Service provides 246 products divided into four main components: ● Global: (65 products) provides a series of biogeophysical products on the status and evolution of the land surface at global scale at mid and low spatial resolution. ● Pan-European: provides information about land	**Parameter**: none found but some free text elements mappable to known vocabularies **Instrument**: none found **Platform**: none found **Generic keywords**: Terms from Eionet, OZCAR, GCMD, INSPIRE, ISO 19115, NVS	**5.1.2** **5.1.3** **5.2.1**

Data sources	Type and number of datasets (from D2.2 and D5.1 deliverables)	Result of high level semantic analysis of DAB harmonised XML	Pilots
	cover and land use and its changes, as well as biogeophysical parameters at European scale at high resolution ● Local: focuses on different hotspots, i.e. areas that are prone to specific environmental challenges and problems ● Imagery and reference data: satellite imagery forms the input for the creation of Land Monitoring products; and in order to ensure the efficient use of satellite imagery, in-situ data is required **Datasets identified by Pilots** ● 5.1.2: Delineation of Riparian Zones; Green Linear Elements; EU-DEM v1.1 is a digital surface model (DSM); Imperviousness; European Settlement Map 2016, 2017, 2019; Corine Land Cover 1990,2000, 2006, 2012, 2018 (100m resolution) ● 5.1.3: Sentinel-1 Single Look Complex products; ECMWF ERA5 hourly data ● 5.2.1: ERA5 provides hourly estimates of a large number of atmospheric, land and oceanic climate variables	**Comment**: ● Dataset titles could be used as a source of additional keywords e.g. Global ocean; real time; in-situ observations; objective analysis; Arctic; ocean colour; plankton; transparency; L4; NRT; monthly observations; monthly mean; sea surface; wind and wind stress; scatterometer; model	

Funded by the European Union

Data sources	Type and number of datasets (from D2.2 and D5.1 deliverables)	Result of high level semantic analysis of DAB harmonised XML	Pilots
NCEI data (US-NODC)	The NCEI Geoportal is different from the NOAA OneStop tool as it seems to be a subset of onestop. You can check this by searching for "NOAA" in onestop and going to collection attributes -> data centres. You will find "National Centers for Environmental Information, NESDIS, NOAA, U.S. Department of Commerce" having about 40,000 records amongst other data centres. These records are the one found in NCEI geoportal. NCEI is however the biggest source **Datasets identified by Pilots** ● 5.2.1: NOAA-WOA / Temperature, Salinite, Percent Oxygen saturation, Nitrate, dissolved oxygen	**Parameter**: Good semantic description at usage level with what seems to be consistent use of P07 URIs and/or GCMD URIs **Instrument**: Good semantic description using GCMD instruments with many already aligned to L05 **Platform**: looks like consistent mapping to C17 for marine observations **Generic keywords**: Some matches to terms from EIONET, OZCAR, GCMD, INSPIRE, ISO 19115, NVS	**5.2.1**
ENA	The European Nucleotide Archive (ENA) provides a comprehensive open record of the world's nucleotide sequencing information and a platform for the management and analysis of sequence and related data. Covering raw sequencing data, sequence assembly information, functional annotation and a host of further data types, content is measured in millions of taxa, hundreds of thousands of sequenced libraries and petabytes of storage. **Datasets identified by Pilots** ● 5.3.1: sequence data	**Parameter**: NCBI taxon; terms form EnvO **Instrument**: Some matches to terms defined in the Experimental Factor Ontology (EFO) **Platform**: no information **Generic keywords**: Terms from EnvO	**5.3.1**
EEA	The EEA Datahub contains 209 datasets from a wide range of different topics, like air pollution, energy,	**Parameter**: some terms from EIONET vocabs in keyword section	**5.1.2**

Data sources	Type and number of datasets (from D2.2 and D5.1 deliverables)	Result of high level semantic analysis of DAB harmonised XML	Pilots
	water and many more. **Datasets identified by Pilots** ● 5.1.2: Water Information System for Europe (WISE); European inventory of nationally designated protected areas; Ecosystem types of Europe; Natura 2000 data - the European network of protected sites;	**Instrument**: none found **Platform**: none found	

5.3 DCAT-FE Turtle example

```
@prefix ex: <http://https://fairease.eu/dataset/> .
@prefix dap: <http://example.org/dap-example/> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@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 ssn: <http://www.w3.org/ns/ssn/> .
@prefix xsd: <http://www.w3.org/2001/XMLSchema#> .
@prefix dcat: <https://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 skos: <http://www.w3.org/2004/02/skos/core> .
@prefix sdo: <https://schema.org/> .


ex:FAIREASECatalog a dcat:Catalog ;
  dct:title "FAIREASECatalog Catalog"@en ;
  rdfs:label "FAIREASECatalog Catalog"@en ;
  foaf:homepage <http://example.org/catalog> ;
  dct:publisher <EDMERP:FAIR-EASE> ;
  dct:language <http://id.loc.gov/vocabulary/iso639-1/en>  ;
dcat:theme <https://vocab.nerc.ac.uk/collection/P22/current/28/> ;
dcat:catalog ex:SDNCatalog, ex:CopernicusCatalog ;
dct:identifier <https://cdi.seadatanet.org/report/681> ;
adms:identifier [
rdf:parseType "Resource";
skos:notation "10.1000/182" ;
];

dct:issued "2008-12-04".
ex:SDNCatalog a dcat:Catalog .
ex:CopernicusCatalog a dcat:Catalog .

ex:SDNCatalog dcat:dataset ex:MyDataset.

ex: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. ";

dcat:distribution ex:MyDataset-001-csv ;
dqv:hasQualityMeasurement ex:measurement1;
dct:publisher  ex:INFRstructureArgo;
dct:creator  <https://edmo.seadatanet.org/report/43> ;


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

dct:publisher <https://www.bodc.ac.uk/>;
prov:wasGeneratedBy dap:Activity1 .


dap:Activity1 a prov:Activity;
    prov:used <http://vocab.nerc.ac.uk/collection/L05/current/134/> .

<http://vocab.nerc.ac.uk/collection/L05/current/134/>  a sosa:Sensor, prov:Entity;
    sosa:observes <http://vocab.nerc.ac.uk/collection/P01/current/TEMPP681/> ;

#or we can break this down to its iadopt properties
 # sosa:observes [
 #  hasObjectOfInterest      <http://vocab.nerc.ac.uk/collection/S21/current/S21S027> ;
 #  hasProperty       <http://vocab.nerc.ac.uk/collection/S06/current/S0600160/> .
#]

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> ;
    ] ;

sosa:isHostedBy <http://vocab.nerc.ac.uk/collection/B76/current/B7600031/> .

<http://vocab.nerc.ac.uk/collection/B76/current/B7600031/> a sosa:Platform;
sosa:hosts <http://vocab.nerc.ac.uk/collection/L05/current/134/> .
<http://vocab.nerc.ac.uk/collection/P01/current/TEMPP681/> a sosa:ObservedProperty .


##Distribution
ex: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 "distribuciÃ³n en CSV del conjunto de datos imaginario 001"@es ;
  dcat:mediaType <http://www.iana.org/assignments/media-types/text/csv> ;
  dcat:accessService ex:subset-service-001 ;
  dcat:byteSize "5120"^^xsd:nonNegativeInteger .


  ex:measurement1 a dqv:QualityMeasurement ;
  dqv:isMeasurementOf  <L27>;
#Check if an L20 term could be applied

    dqv:value "good"^^xsd:boolean .

ex: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 ex:MyDataset .
```

Funded by the European Union