D9.6
Marine subdomain EOV product specification

Work Package	WP9
Lead partner	IFREMER
Status	Final
Deliverable type	Report
Dissemination level	Public
Due date	30-04-2020
Submission date	11-05-2020

Deliverable abstract

The overarching goal of ENVRI-FAIR is that all participating research infrastructures (RIs) will improve their FAIRness and become ready for connection of their data repositories and services to the European Open Science Cloud (EOSC). Deliverable 9.1 has reported on the roadmap of the RIs in the marine subdomain towards improving their FAIRness. It presented the approach of using FAIR questionnaires (together with WP5) to identify the strengths and weaknesses of each RI and a first indicative set of activities to improve identified weaknesses or gaps. After formulation in Deliverable D9.2 of implementation plans for mitigating these gaps during the next phase of the ENVRI-FAIR project, the RIs from the marine subdomain have specified in Deliverable D9.3 the technical services and interfaces to be implemented at RI level and have undertaken the implementation. The RI services will be demonstrated in D9.4 (M27) and will be operational for EOSC operations in D9.5 (M36). The present deliverable D9.6 is linked to Task 9.8 going from M24 to M48 and aiming to demonstrate the marine subdomain FAIRness. Here the This document D9.6 describes the technical specifications that will be necessary for the Marine EOV product Version 1 and Version 2 to run in 2021 and 2022.

ENVRI-FAIR (www.envri-fair.eu) has received funding from the European Union’s Horizon 2020 research and innovation programme under grant agreement No 824068

DELIVERY SLIP

	Name	Partner Organization	Date
Main Author	Thierry Carval	IFREMER	22-02-2021 18-03-2021
Contributing Authors	Peter Thijsse Marc Portier Katrina Exter Benjamin Pfeil Sylvie Pouliquen Ivan Rodero Alexandra Kokkinaki Justin Buck Antoine Quéric Valérie Harscoat	MARIS Lifewatch (VLIZ) Lifewatch (VLIZ) UIB Euro-ARGO EMSO ERIC BODC BODC Euro-ARGO IFREMER	18-03-2021
Reviewer(s)	Angeliki Adamaki Damien Boulanger	ULUND CNRS (IAGOS)	19-04-2021
Approver	Andreas Petzold	FZJ	11-05-2021

DELIVERY LOG

Issue	Date	Comment	Author
V0.1	22-02-2021	Draft version submitted to partners prior to the specification meeting on 24th Feb 2021	T. Carval
V0.2	From 24-02-2021 to 18-03-2021	Suggestions and comments of RIs	All the contributors
V0.3	18-03-2021	Internal review version (processing all suggestions and comments of RIs)	T. Carval
V0.4	19-04-2021	Internal review	A. Adamaki, D. Boulanger
V0.5	21-04-2021	Processing review feedbacks + layout revision + References + Glossary	V. Harscoat
V1.0	28-04-2021	Finalised version	T. Carval & all contributors

DOCUMENT AMENDMENT PROCEDURE

Amendments, comments and suggestions should be sent to the Project Manager at manager@envri-fair.eu.

GLOSSARY

A relevant project glossary is included in Appendix A. The latest version of the master list of the glossary is available at http://doi.org/10.5281/zenodo.4471374.

PROJECT SUMMARY

ENVRI-FAIR is the connection of the ESFRI Cluster of Environmental Research Infrastructures (ENVRI) to the European Open Science Cloud (EOSC). Participating research infrastructures (RI) of the environmental domain cover the subdomains Atmosphere, Marine, Solid Earth and Biodiversity / Ecosystems and thus the Earth system in its full complexity.

The overarching goal is that at the end of the proposed project, all participating RIs have built a set of FAIR data services which enhances the efficiency and productivity of researchers, supports innovation, enables data- and knowledge-based decisions and connects the ENVRI Cluster to the EOSC.

This goal is reached by: (1) well defined community policies and standards on all steps of the data life cycle, aligned with the wider European policies, as well as with international developments; (2) each participating RI will have sustainable, transparent and auditable data services, for each step of data life cycle, compliant to the FAIR principles. (3) the focus of the proposed work is put on the implementation of prototypes for testing pre-production services at each RI; the catalogue of prepared services is defined for each RI independently, depending on the maturity of the involved RIs; (4) the complete set of thematic data services and tools provided by the ENVRI cluster is exposed under the EOSC catalogue of services.

TABLE OF CONTENTS

D9.6 - Marine subdomain EOV product specification

1 Rationale

The Task 9.8 "Marine Essential Ocean Variable - EOV" (started end of 2020) is a use case based on the data and metadata services set up by the Research Infrastructures – RIs. It will provide interoperable access to RI data to end users, in particular the VIP users CMEMS, SeaDataNet or EMODnet involved in the project.

Each RI involved in WP9 has developed an implementation plan [D9.2] that addresses the results of their FAIRness self-analysis, with the shared objective to improve the FAIRness at the Marine subdomain level as illustrated in the following figure.

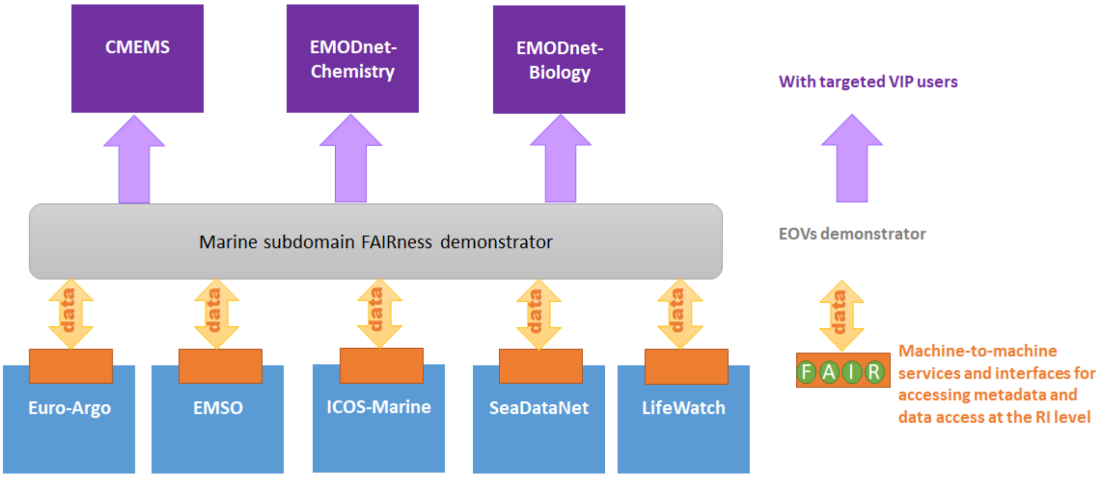
CMEMS
EMODnet-Chemistry
EMODnet-Biology
With targeted VIP users
Marine subdomain FAIRness demonstrator
EOVs demonstrator
data
Euro-Argo
EMSO
ICOS-Marine
SeaDataNet
LifeWatch
F A I R
Machine-to-machine services and interfaces for accessing metadata and data access at the RI level

Figure 1: Marine sub-domain implementation plan

The demonstrator targeted within Task 9.8 aims to set up a workflow that will serve data files answering to EOV requests, in particular of the VIP users (SeaDataNet, CMEMS and EMODnet) to allow them to process data aggregation and product assessment (QA/QC) on the extracted data.

This document D9.6 describes the technical specifications that will be necessary for D9.7 (the Marine EOV product to run in 2021) and D9.8 (to run in 2022). The demonstration will be implemented with a Jupyter notebook: a search will be made on the data/metadata services of the RIs on a specific set of EOVs (oxygen or chlorophyll-A, temperature, salinity, and zooplankton biomass and diversity). Here we describe the (meta)data that need to be provided, the format and ways in which the (meta)data should be provided, outline the general architecture of the broker services that will perform the exercise, and describe the details that are specific to each RI. Finally, we discuss what changes are currently envisaged for D9.8; as this deliverable will run in 2022, it will be able to benefit from the ongoing Task Forces activities.

2 General specification

2.1 Context

With the “Marine EOV” use case, WP9 partners will be among the users of the enhanced Marine data services developed within ENVRI-FAIR.
The “Marine EOV” solutions (the broker) are targeted toward major stakeholders: EOSC, CMEMS, EMODnet, SeaDataNet. We assume that they are relevant for new initiatives such as Blue-Cloud, EuroSea, the Digital Twin of the Ocean.

2.2 Requirements

An example of user request is: “I want to access the data (as exposed by the RIs) having oxygen EOV observed in 2020 in Arctic from the Marine RIs”

This request is possible when RIs expose data with interoperable vocabularies and provide the dates and locations metadata.

The data that are provided must include the EOV parameters and the temporal and geographic scope and must also make use of interoperable vocabularies, in particular those from BODC NVS.
The types of data that are provided, however, is for each RI to decide: individual scientist/instrument-created data files, aggregated data products, computed data etc., are all within the scope, with the main requirement that they must be accessible via ERDDAP for the 2021 scenario.

Note that no AAI step will be involved in the exercise, therefore it is assumed – but not per se required – that the data being accessed are open access.

2.3 Architecture

The main components are:

- A broker to orchestrate the request
- A vocabulary server to manage the EOVs – RIs parameter descriptions
- For each RI
    - A metadata API to query RI parameters
    - A data API
    - A formatting service to generate files from data and metadata
        - ERDDAP is the initial solution for 2021 release, and this will combine the metadata and data API into a single API (V1)
        - SPARQL endpoints will be an option for 2022 release (V2)
- For each integrator (non-ENVRI-FAIR user)
    - A data scientist to assess the aggregated data (i.e. the heterogeneous set of datafiles which are assembled as a result of the EOV requests to each RI), to provide feedback on anomalies by way of RI-unique PIDs
    - Download logs from integrators for RI statistics (cf. AtlantOS Semaphore)

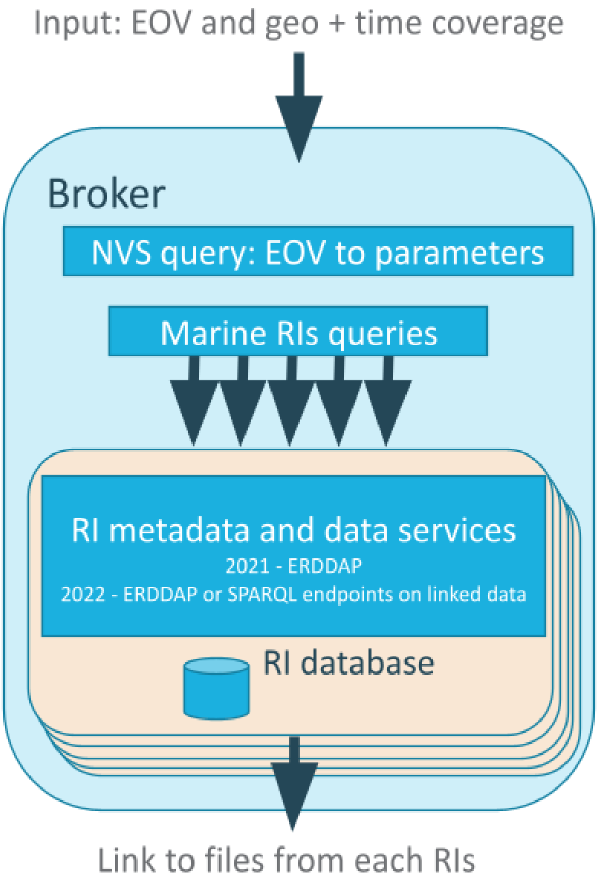
Input: EOV and geo + time coverage
Broker
NVS query: EOV to parameters
Marine RIs queries
RI metadata and data services
2021 - ERDDAP
2022 - ERDDAP or SPARQL endpoints on linked data
RI database
Link to files from each RIs

Figure 2: the main components of Marine EOV demonstrator

The broker is a python script in a Jupyter NoteBook. It can be reused and showcased in different ways: from a script, from a User Interface, from any 3rd party integration work.
A series of libraries will be evaluated with the broker such as:

- https://github.com/ioos/erddapy, a python library to interact with ERDDAP via Jupyter
- https://intake.readthedocs.io/en/latest/ intake is a python library (generic) that helps one to access data from a dedicated catalogue. Developing new interfaces is straightforward; intake could build ERDDAP queries for a user based on what they provide as arguments

The broker can be operated by any user interested in Marine multiple RI data access from a Jupyter Notebook, that can be activated on any server having an internet link to NVS and RI data services (ERDDAP, SPARQL endpoints).

There is no specific GUI (graphic user interface); any user or data portal can configure its own GUI. Ifremer, EMSO and others will provide GUI examples such as libraries made available to Jupyter NoteBook users. EMSO MOODA is providing this type of functionality and it could integrate the broker within the MOODA library

- EMSO mooda modules and widgets (https://rbardaji.github.io/mooda/, https://rbardaji.github.io/mooda/docs/examples/emso-qc-widget.html)

ERDDAP main features are:

- An example : EMSO PAP site data https://linkedsystems.uk/erddap/tabledap/ENVRIplus_b122nnnn.html
- Can produce a query for the an EOV, for example “salinity”, and it will return data via a URL https://linkedsystems.uk/erddap/tabledap/ENVRIplus_b122nnnn.ncCF?depth%2Clongitude%2Clatitude%2CPOSITION_SEADATANET_QC%2Ctime%2CTIME_SEADATANET_QC%

2CPSALPR01%2CPSALPR01_SEADATANET_QC&time%3E=2002-10-06T19%3A59%3A59Z&time%3C=2007-08-03T00%3A00%3A00Z

- Dataset is based on SeaDataNet NetCDF
  - CF compliant and ACDD https://wiki.esipfed.org/Attribute_Convention_for_Data_Discovery_1-3
  - Also, ERDDAP metadata is based on DCAT so schema.org follows and can see metadata at https://datasetsearch.research.google.com/search?query=bodc%20erddap&docid=xntBQCKNaqaqo7juAAAAAA%3D%3D
- Minimum metadata
  - Positional data and NVS names for variables
  - Add in DCAT https://www.w3.org/2016/11/sdsvoc/SDSVoc16_paper_30 for metadata to underpin 2022 ambitions
- A good example of a visual interface built on top of ERDDAP is the MI digital ocean https://www.digitalocean.ie/

3 Release V1 -2021

3.1 Scenario

Figure 3 lists the main resources involved in this release of the Marine EOV use case.

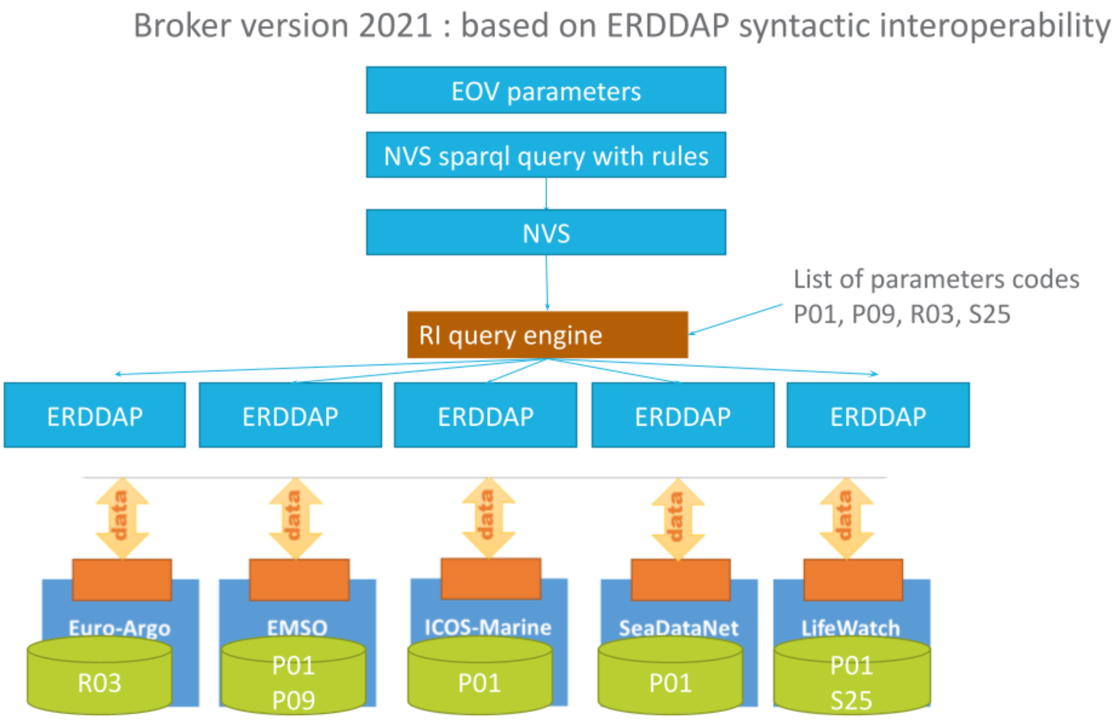

Figure 3: data and metadata resources involved in Marine EOV – revision 1

The scenario for data request is as follows

1. A request is submitted to the broker
2. The broker parses the request to NVS (NERC Vocabulary Server)
3. NVS returns a list of acceptable P01 (parameters) codes for the selected EOV
4. The broker queries all Marine domain data APIs (ERDDAP)

5. The broker aggregates the links to data provided by APIs in a result list
6. The result list is forwarded to the requester

3.2 Broker specification

Input parameters

- Temporal coverage
- Spatial coverage
- List of requested EOVs

Check parameters

- The temporal coverage is valid
- The spatial coverage is valid
- Each EOV is valid

Convert EOV list into RI - Parameters lists

For each RI

- The broker queries the NVS with RI parameter code table and the EOV list
- The NVS returns the parameters list corresponding to the EOV list

Query RI data

For each RI

- The broker generates the data query request based on the above query result of parameters and/or spatiotemporal variables

Output

- The queried data and metadata are returned.
- As part of the metadata provided with the ERDDAP output, it is expected that the following will be provided: dataset usage license, link to the metadata catalogue entry (or similar) for the dataset to provide the provenance trail.
- The data will be formatted by individual ERDDAP servers.

3.3 RI data and metadata services

3.3.1 NVS vocabulary server

The NERC Vocabulary Server provides the parameter tables used by RIs and the relations between EOVs and the RIs parameter codes.
https://vocab.nerc.ac.uk/
Technical contact: AK (BODC)

3.3.2 RI vocabularies and ERDDAP servers

3.3.2.1 Argo

Argo parameters vocabulary is R03 (https://vocab.nerc.ac.uk/collection/R03/current/) .
https://www.ifremer.fr/erddap/tabledap/ArgoFloats.html
Technical contact: AQ (Ifremer)

3.3.2.2 EMSO

EMSO parameters vocabularies are P01 or P09.
http://erddap.emso.eu
Technical contact: IR (EMSO)

3.3.2.3 ICOS-Marine

ICOS-Marine parameters vocabulary is P01.
https://erddap.icos-cp.eu
Technical contact: SJ and JP (UIB)

3.3.2.4 SeaDataNet

SeaDataNet parameters vocabulary is P01.
https://www.ifremer.fr/erddap
Technical contact: AQ (Ifremer)

3.3.2.5 LifeWatch

The LifeWatch data that will be provided will be the LifeWatch datasets in the EurOBIS collection. These data are published in EurOBIS in DwC-A format (via IPT). Within this data format, the parameters are linked to vocabularies and the geographical and temporal scope are provided (also in standardised format). In detail

- Parameters vocabulary is (mostly) P01.
    - P01 covers the EOVs oxygen or chlorophyll-A, temperature, salinity, and the biological EOV zooplankton biomass and diversity, which is linked to P01 through A05's EV-ZOO (http://vocab.nerc.ac.uk/collection/A05/current/EV_ZOO/)
    - However, our biological datasets that will provide the latter EOV have additional columns that provide the actual taxonomic data. This is in fact in line with the usage context of these P01 terms declaring 'the biological entity to be specified elsewhere'. However, for these extra columns the terms are not (fully) covered by BODC but rather by other standard vocabularies, for example
        - Name of the species → http://vocab.nerc.ac.uk/collection/P01/current/SCNAME01/
        - ID of the species → http://vocab.nerc.ac.uk/collection/P01/current/SNANID01/
        - and similar for classification, traits, sex, ……
    - For the biologist, these taxonomic data are of prime importance – almost *all* data searches conducted by biologists will include a taxonomic term. To be unable to search on these terms (by re-using the demonstrator code in the Jupyter notebook) within our data as provided via ERDDAP, would be a backwards step in our data provision. This is something we wish to address alongside the 2021 exercise, potentially to become part of the 2022 exercise.
- Temporal scope is in ISO standard
    - Depicted by term → http://purl.org/dc/terms/modified
- Geographical scope is given in longitude and latitude, as well as through a locationID identifier (if present)
    - Lat → http://rs.tdwg.org/dwc/terms/decimalLatitude
    - Lon → http://rs.tdwg.org/dwc/terms/decimalLongitude
    - Location → http://rs.tdwg.org/dwc/terms/locationID
- The dataset licence is provided in one of the XML files that constitute the DwC-A format (specifically, this is part of the metadata record, not the data)
- Additional information are available covering some of the provenance / methodology (using BODC's Q01 collection)

Technical contact: MP (VLIZ/Lifewatch Belgium)

4 Release V2 - 2022

4.1 Scenario and Broker update

The broker will be enhanced to query RIs having a SPARQL endpoint in 2022. Note that there is no obligation for a RI to implement a SPARQL endpoint.

In the release (see Figure 4 below):

- ERDDAP will provide direct access to data and data sub-setting using parameter queries.
- The SPARQL endpoints will provide semantic queries to a wealth of metadata and links to data.

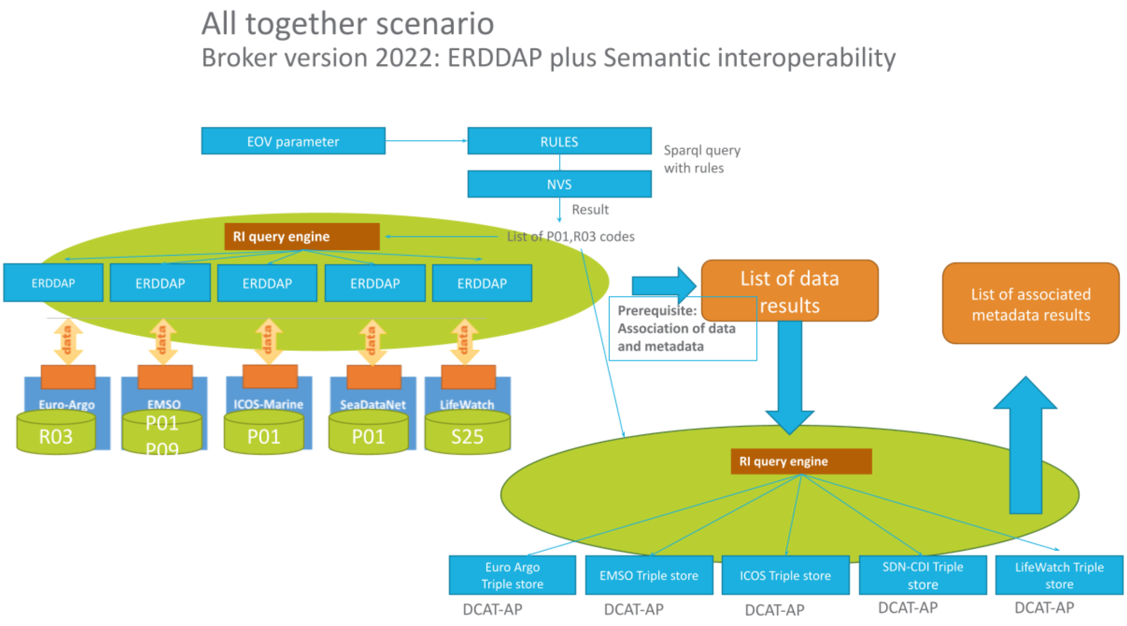

Figure 4: data and metadata resources involved in Marine EOV – revision 2

Some Marine RIs have SPARQL endpoints up and running, some are under development or planned. The same is true for non-marine environmental subdomains.

The RIs having an operational SPARQL endpoint will have three options for 2022 release:

- **Option 1:** the RI SPARQL endpoint will provide additional metadata to the existing ERDDAP data queries.
- **Option 2:** the RI SPARQL endpoint will provide metadata and data. The ERDDAP server will no longer be queried by the broker. Data may be provided by FAIR services such as OGC.
- **Option 3:** a RI may decide depending on datasets to provide option 1 and option 2.

4.2 Task forces feedback

The 6 task forces (TF1 to TF6) are agile activities on RIs common requirements. They organize testbeds for metadata services, community standards and cataloguing.

- TF1 Catalogue of services
- TF2 ENVRI AAI implementation
- TF3 PIDs identification types and registries
- TF4 Triplestores, vocabularies - provenance
- TF5 Licenses citation and usage tracking
- TF6 ENVRI-Hub

There is work underway within the Task Forces that will impact the Marine EOV use case. The Marine EOV 2022 release will implement outcomes from the Task Forces. The Marine RIs are well represented in the aforementioned activities, with special focus on the following groups.

4.2.1 TF6: cross domain data queries

TF6 will develop a cross domain use case similar to Marine EOV.

4.2.2 TF4: SPARQL endpoints, provenance information

TF4 is supporting triple store and SPARQL endpoint implementation.
Provenance information should be provided with the data objects that lie at the end of the deliverables, and at least a minimum set of such information should be defined. This should include what provenance information is required, how that information should/could be provided, and which vocabularies those information should/could be expressed in. This is work that is best done together with the task forces, in particular TF4 and TF3 as these data and the WP9 exercise provides a concrete use-case for their outputs. Our initial ideas for categories of provenance information include

- Data creator (person information, institute information, etc)
- Data collection site (cruise, platform, etc)
- Instrumentation
- Sampling/data processing information (methods, protocols, standard operating procedures, software, etc)
- Quality control information (methods, protocols, standard operating procedures, software, data level [raw, processed, etc], etc)
- Links to related publications (DOIs, URLs, etc)

4.2.3 TF5: licence information

The “Marine EOV” queries multiple datasets from multiple RIs having distinct data licences.
With TF5 support, we shall properly expose the datasets licences.

5 References

Ref	Title	Version / Date
D9.1	Marine subdomain FAIRness roadmap https://iagos-comm.iek.fz-juelich.de/dmsf/files/3946/view	V2.0 August 2019
D9.2	Marine subdomain implementation plan https://iagos-comm.iek.fz-juelich.de/dmsf/files/3944/view	V1.0 November 2019
D9.3	RIs technical specification https://iagos-comm.iek.fz-juelich.de/dmsf/files/4182/view	V1.0 18 March 2020

6 Appendix A

AAI	Authentication and Authorisation Infrastructure
ACDD	Attribute Convention for Data Discovery
API	Application Programming Interface
CDI	Common Data Index (metadata format and data access system by SeaDataNet)
CF	Climate and Forecast (semantics for NetCDF)
CMEMS	Copernicus Marine Environment Monitoring Service
COPERNICUS	A major earth observation programme run by European Commission and European Space Agency
CP	Carbon Portal
DwC-A	Darwin Core Archive file format
EMSO	European Multidisciplinary Seafloor and water column Observatory
ENVRI	1) An environmental RI cluster FP7 project 2) Environment research infrastructures (in ESFRI level or upcoming) as a community
EOSC	European Open Science Cloud
EOV	Essential Ocean Variable(s)
ERDDAP	NOAA developed science data server technology
ERIC	European Research Infrastructure Consortium (legal entity type)
ESFRI	European Strategy Forum on Research Infrastructures
EuOBIS	European OBIS
FAIR	Findable Accessible Interoperable Reusable
GUI	Graphical User Interface
ICOS	Integrated Carbon Observation System
IPT	Integrated Publishing Toolkit
M	Month
NetCDF	Network Common Data Format
NVS	NERC Vocabulary Serveur
NOAA	US National Oceanic and Atmospheric Administration
OBIS	Ocean Biogeographic Information System
OGC	Open Geospatial Consortium
PID	Persistent Identifiers
QA/QC	Quality Assurance/Quality Control
RDF	Resource Description Framework
RI	Research Infrastructure
SDN	SeaDataNet pan-European infrastructure for marine data management
SPARQL	SparQL Protocol and RDF Query Language
TF	Task Force