D9.2
Marine subdomain implementation plan

Work Package	WP9
Lead partner	MARIS
Status	Final Draft
Deliverable type	Report
Dissemination level	Public
Due date	30-09-2019
Submission date	30-11-2019

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. In this Deliverable 9.2 the RIs in the marine subdomain have formulated implementation plans for mitigating these gaps during the next phase of the ENVRI-FAIR project.

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	Peter Thijsse	MARIS	30-08-2019 24-10-2019 14-11-2019 20-11-2019
Contributing Authors	Dick Schaap Katrina Exter Alex Vermeulen Benjamin Pfeil Thierry Carval Sylvie Pouliquen Ivan Rodero	MARIS VLIZ ICOS ERIC UIB IFREMER IFREMER EMSO ERIC	30-08-2019, 19-11-2019 14-11-2019 18-11-2019 18-11-2019 18-11-2019 18-11-2019 14-11-2019
Reviewer(s)	Daniela Franz	FZJ	26-11-2019
Approver	Andreas Petzold	FZJ	30-11-2019

DELIVERY LOG

Issue	Date	Comment	Author
V 0.1	30-08-2019	Approach and set-up of Deliverable 9.2	Peter Thijsse and Dick Schaap
V0.2	09-09-2019	Reviewed approach	Sylvie Pouliquen, Dick Schaap
V0.3	24-10-2019	Initial version with requests for more input	Peter Thijsse
V0.4	14-11-2019	Complete draft to RIs for review	Peter Thijsse
V0.5	19-11-2019	Internal review	Dick Schaap
V1.0	20-11-2019	Finalised draft processing all suggestions and comments of RIs	Peter Thijsse

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.3465753.

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

1. Introduction

The ENVRI-FAIR project is engaging Research Infrastructures (RIs) in the environmental domain covering the subdomains Atmosphere, Marine, Solid Earth and Biodiversity / Ecosystems. 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).

WP9 has a focus on the RIs in the Marine subdomain, which are represented in ENVRI-FAIR by Euro-Argo, EMSO, and the marine component of ICOS and LifeWatch, as RIs listed on the ESFRI roadmap, as well as SeaDataNet as European marine data management infrastructure. The overall aim of WP9 is to analyse the FAIRness of each of these RIs and to implement within each RI the necessary actions to improve its FAIRness. The latter is critical for the Marine subdomain as it will provide a coherent base for developing the integrated services systems required by a broad variety of research, regulatory and operational communities.

In Deliverable 9.1 the RIs in the marine subdomain have reported on the roadmap towards improving their FAIRness. This document 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. In this Deliverable 9.2 the RIs in the marine subdomain reflect further on their FAIRness gaps and present implementation plans to mitigate these gaps during the next phase of the ENVRIFAIR project.

2. Approach

As mentioned above the basis for this deliverable is the finalized Deliverable D9.1 - Marine subdomain FAIRness roadmap that is downloadable from REDMINE: https://iagos-comm.iek.fz-juelich.de/dmsf/files/3709/view.
D9.1 is the result of Task 9.2 "Analysis and priorities to enhance RI data FAIRness" and it has been compiled with contributions of WP9 partners, representing the RIs within the marine subdomain, while interacting with WP5 (Community standards and catalogue of services) and WP7 (Common implementation and support).
D9.1 gives an analysis and assessment of the FAIRness of each of the marine RIs, resulting in Chapter 5 in an overview per RI of current strengths, identified gaps and planned and/or suggested directions for solutions to overcome those gaps. The next challenge for Task 9.2 has been to work out conceptual implementation plans for each marine RI, based on the findings in Chapter 5 of D9.1. This analysis was performed by each of the marine RIs, resulting in an implementation plan per RI. These have been bundled in this Deliverable 9.2 - Marine subdomain implementation plan.

After discussion with WP5 and WP7, the following approach has been agreed and followed to extract the ideas for the marine subdomain implementation plan and to provide input for follow-up activities as part of WP5 and WP7.

Step 1
Each of the marine RIs has worked out a conceptual implementation plan for their RI. While doing this, it has been considered that it is not possible to improve all identified gaps / weaknesses within the framework of ENVRI-FAIR. Therefore, in some cases a subset has been determined, which will actually be worked out in the ENVRI-FAIR project.
Each RI has been asked to undertake analysis and report, using a simple template as follows:

- Summarise earlier identified gaps / weaknesses
- Prioritise identified gaps / weaknesses and make a selection which ones will be worked out as part of ENVRIFAIR. This evaluation should be done and described considering:

1. Already known user requirements at RI level;
2. What is required at marine subdomain level to facilitate Task 9.8 "Demonstrate marine subdomain FAIRness, EOV global product".

- Per selected gap / weakness: suggested approach for improving;
  - Technologies to be adopted/developed; could still be options and no firm choices. This should be at conceptual technical level;
  - Planning in time;
  - Involved partners (if relevant).

After this the commonality of the individual plans has been initially analysed at the cluster level.

Step 2
First evaluation of the draft implementation plans by WP5 and WP7 teams:

1. WP5 has evaluated the plans from the perspective of common ENVRI-FAIR standards, where possible or relevant.
2. WP7 has identified a first set of possible common issues between RIs and associated solution options. For that purpose, WP7 has set-up and is populating a Knowledge Base (KB) of technical solutions and best practices for implementing these, partly by querying the internet, partly by trying it out themselves, and partly by asking and receiving input from ENVRI-FAIR RIs.

A first review of the draft implementation plans of RIs in ENVRI-FAIR by WP5 and WP7 was carried out during the WP5 review meeting in Lund (29/10/2019). An important outcome of that meeting was a global analysis of cross-cutting gaps/themes. Initially it was planned to go a step further in the feedback and in-depth analysis of the RI implementation plans, but during the meeting it was decided to limit to a global review, first feedback and setting up cross-cutting task forces. This decision was made due to the time-limit towards the deadline, and due to the fact that other subdomains than the marine subdomain (with different time paths) did not have their draft implementation plans yet. Therefore, an in-depth review of the technical solutions could not be done on-time for D9.2, but will be done early 2020. Results of this action will then be included in D9.3: the RI's technical specifications.
The taskforces will stimulate discussions and collect important feedback on the following themes:

- ENVRI (VO) AAI implementation
- PID's, identification, types and registries
- Triple stores, data storage, certification
- Licenses, citation, usage tracking
- User feedback and demonstration cases in Jupyter
- ENVRI catalogue of services

These themes were extracted from the WP9.2 draft contributions created by the RI's as preparation for the Lund workshop. Each taskforce will consist of one or more WP7 specialists, a WP4/5 specialist, plus at least one participant per interested RI. All taskforces will focus on APIs and metadata that are the glue in FAIRness. All RI/Subdomain participants that will be part of the taskforces will participate with their subdomain use case in mind, as this is the key priority for them to use certain solutions.

Step 3
Each of the marine RIs has finalised their implementation plan to be integrated in D9.2.

1. **Only considering** the first suggestions and recommendations as received from WP9 internally.
2. Taking into account as explained above, that the next Deliverable D9.3 by March 2020 will include more detailed technical choices, priorities in development, and planning, and that this deliverable will include input and detailed feedback from the WP5/WP7 task teams.

3. Implementation plans

3.1 ICOS-marine

3.1.1 Overview

ICOS provides consistent, long-term and high-quality observations required to understand the present state and predict future behaviour of the global carbon cycle and greenhouse gas emissions. ICOS RI is a distributed research infrastructure where central facilities (CFs) perform initial quality control, provide second level QCed data from the network and are responsible for the metadata and file content while the ICOS Carbon Portal is responsible for the technical infrastructure of long-term archival and making it available (see Figure 1). In respect to achieving FAIRness within WP9, ICOS-marine (University of Bergen) as the CF managing the marine part of ICOS is responsible for achieving FAIRness of the metadata and data content FAIR, while the ICOS Carbon Portal (University of Lund) is responsible for the technical implementation and infrastructure.

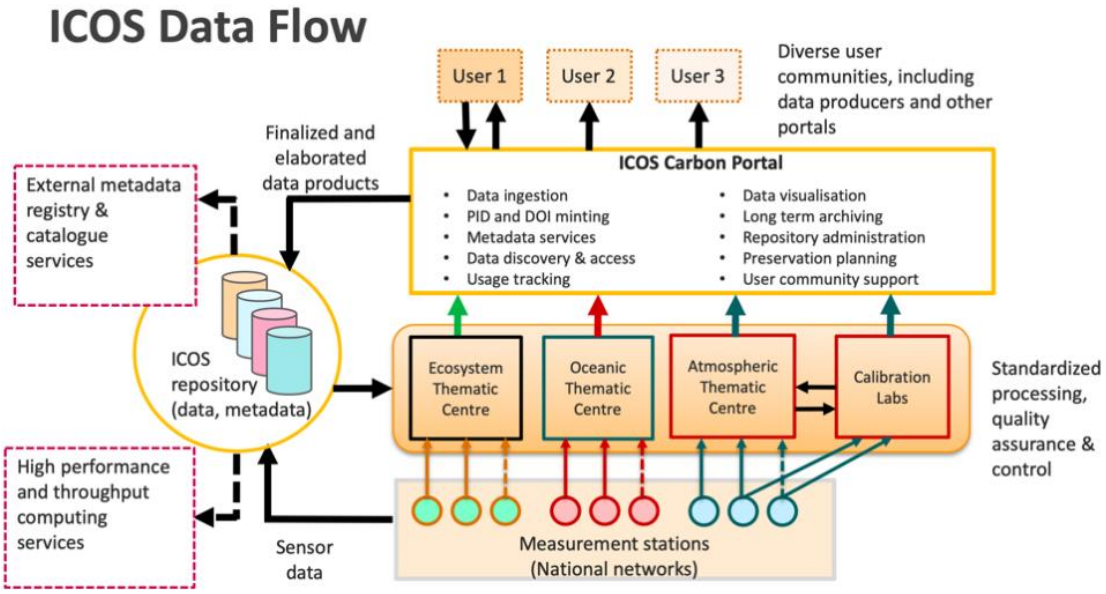

Figure 1: Scheme of the data flow with ICOS RI.

3.1.2 Gaps

ICOS-marine identified the following FAIRness gaps (see also D9.1).

Findability

- Missing support for key words in metadata (present for DOIs)
- Missing plain text search
- Missing schema.org tags in landing pages
- No support for OAI-PMH metadata exchange
- Data collections are not visible in the search portal
- No/limited possibility to find "similar" datasets (based on a given PID/DOI)

Accessibility

- Data license fixed to project
- No standard way to access data object from PID
- No/limited documentation on different ways to access data via machine-based workflows and/or (jupyter) scripts run on ICOS-external servers

Interoperability

- Schemas not registered in common registry
- Controlled vocabularies not completely developed, and published
- Existing standardized vocabulary (UKRI/ NOC and CF) terms are not implemented
- Missing support for subsetting of data
- No/limited support for exporting data & metadata in other formats than ICOS sub-domain standards (ascii/text or netCDF)

Reusability

- Provenance of data processing only in text files and articles
- Reference to machine operable provenance in metadata is often missing
- Linkages (in OpenAire style) between digital objects are not expressed in machine-actionable ways
- Metadata reporting is currently not compliant to the newly established methodologies e.g. United Nations Sustainable Development Goal Methodology 14.3.1

3.1.3 Implementation actions

For each of the gaps ICOS-marine foresees the following actions to improve FAIRness as part of the ENVRI-FAIR project.

Findability

- **Missing support for key words in metadata (present for DOIs)**
  - Explicit support for key words following the Dewey scheme will be added to all object specifications. Object specifications tied to a specific ICOS domain will require that the respective Thematic Centre (TC) provides the correct key words.
  - Next to the key words linked to the object specification each data object can be assigned additional free text key words. This is especially useful for elaborated products.
- **Missing plain text search**
  - Free text search will be added to the search portal and its sparql endpoint. It will search for key words, variable names and description fields.
- **Missing schema.org tags in landing pages**
  - Relevant metadata fields in the html data object landing pages will be tagged with schema.org fields to support Google Data Search.
- **No support for OAI-PMH metadata exchange**
  - A simple service will be developed that maps the sparql output listing all data objects onto the OAI-PMH XML format
- **Collections are not visible in the search portal**
  - Discover the collections based on the data object type in the collection in the search results and have them on top of the find list. Collections can then be expanded into their parts. Support the data objects that are part of a collection using the IsPartOf property, also on the landing pages.
- **No/limited possibility to find "similar" datasets (based on a given PID/DOI)**
  - Implementation of a routine that generates a similarity score based on the comparison of the metadata for a data object defined by the given PID and the other data objects, based on scores based on data object spec, domain, project, variable names, key words and station.

Accessibility

- **Data license fixed to project**
  - All data objects will have an optional field that allows to define a machine actionable data license bypassing the default defined by the project
- **No standard way to access data object from PID**

    - Encode the data link following the recommendation by Starr et al. Add the data link to the Handle registry PID kernel information as soon as this is standardized.
- **No/limited documentation on different ways to access data via machine-based workflows and/or (Jupiter) scripts run on ICOS-external servers**
    - Improve, create and update the documentation on the access to data for automated workflows

Interoperability

- **Schemas not registered in common registry**
    - Register the ICOS schemas in schema.org
- **Controlled vocabularies not completely developed and published**
    - Develop the vocabularies into controlled vocabularies applying the relevant standards and publish them in the relevant registry, this is particularly work for the ICOS domain TCs.
    - For the marine part of the ESFRI landscape standardised controlled vocabularies are already used within certain communities (e.g. European marine data infrastructures/CMEMS) - ICOS will apply those vocabulary lists e.g. UKRI/ NOC and CF vocabulary lists. In addition, other apply metadata standards will be implemented e.g. EDMO, ORCID and for assuring re-usability the enriched metadata as defined by the SDG 14.3.1 methodology for ocean acidification.
- **Missing support for subsetting of data**
    - Setup the APIs for data access and subsetting in agreement with the EOSC and ENVRIFAIR developing standards. As a first step revive the THREDDS server and connect this to the ICOS repository. Take care of the data usage count when subsetting or THREDDS access takes place.
- **No/limited support for exporting data & metadata in other formats than ICOS domain standards (ascii/text or netCDF)**
    - As a first step evaluate the user needs for different data formats and select the data format(s) where needs are highest and urgent. Implement these as possible export format. Note that this breaks the 'do not touch the original data objects' rule and should be clearly communicated to the user.

Reusability

- **Provenance of data processing only in text files and articles**
    - Develop the provenance metadata at TC level using the templates developed in ENVRIplus and expose this through CP
- **Reference to machine operable provenance in metadata often missing**
    - See previous point. Add optional metadata element OtherInfo to data object metadata that links to measurement protocols and articles.
- **Linkages (in OpenAire style) between digital objects are not expressed in machine-actionable ways**
    - No solution identified yet.
- **Metadata reporting is currently not compliant to the newly established methodologies e.g. United Nations Sustainable Development Goal Methodology 14.3.1**
    - Implementation of metadata schemes compliant with UN SDG methodology 14.3.1

Foreseen deadlines for delivery:

- Month (M)18: Content and metadata compliant at OTC
- M27: Implementation in demonstration mode
- M36: Implementation with ICOS RI services

3.2 LifeWatch

3.2.1 Overview

Participanting as part of the LifeWatch RI are LifeWatch consortium and VLIZ. Data systems involved are hosted at VLIZ: EurOBIS, EMODnet, WoRMS, and Marine Regions being the main players, plus MDA and IMIS teams, all based at VLIZ.
The focus for FAIRness improvements will be on the back-end and on metadata and data layer, as illustrated in the diagram below. (Human) User interfaces and everything focused on the user experience will be out of scope, see Figure 2.

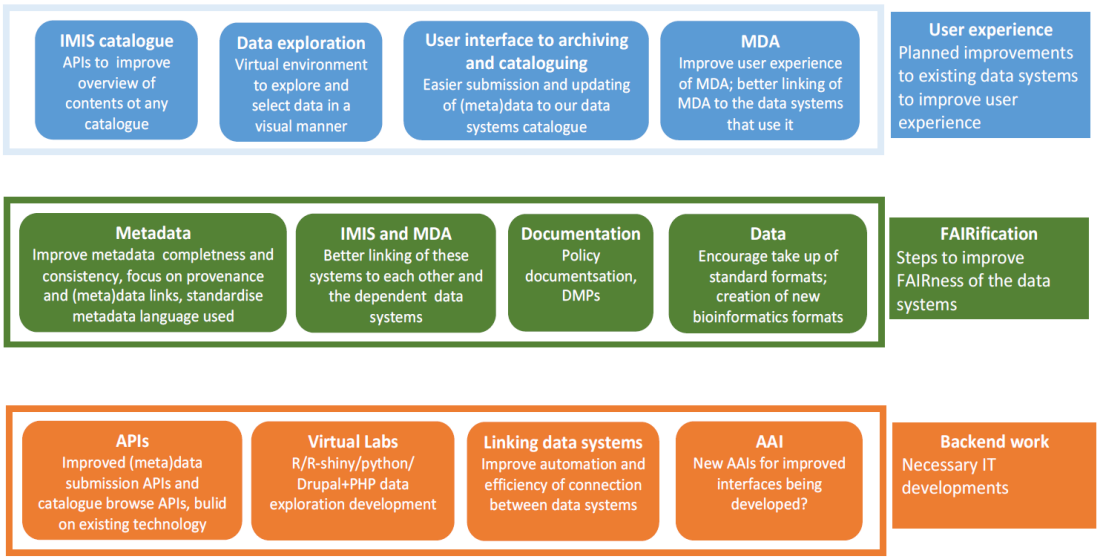

Figure 2: Focus of FAIRness actions for LifeWatch.

3.2.2 Gaps

LifeWatch has considered the following gaps identified in report D9.1.

Findability

- PIDs are not provided for data in the MDA
- PID to the metadata are not included in the metadata in the MDA/EurOBIS/IMIS
- PID to the data are not always included in the metadata in IMIS

Accessibility

- Standardisation of discovery metadata not uniform over entire catalogue (IMIS) or in MDA
- MDA and IMIS not registered
- No explicit policy documents

Interoperability

- No registration of metadata schema in common registry
- No online explanation of metadata
- Not all data are in standard formats

Re-usability

- Limited provenance information
- No DMP for MDA

- Re-use metadata not always consistent within metadata records and between IMIS and MDA, and with the EurOBIS and LifeWatch systems

3.2.3 Implementation actions

Taking into account LifeWatch planning for improvements of the data systems, the following implementation plan has been drafted as part of ENVRI-FAIR. This addresses for each element of FAIR the proposed approaches for bridging the gaps, technologies to be adopted or developed, and initial estimates of timelines (noting that these lean towards the optimistic).

Findability

- **PIDs are not provided for data in the MDA**
- **PID to the metadata are not included in the metadata in the MDA/EurOBIS/IMIS**
- **PID to the data are not always included in the metadata in IMIS**
    - Approach: Metadata does have a PID (the end of the URL is the "IMIS id") which are unique and persistent, however LifeWatch will
        - Add the full PID of the metadata record and of the linked dataset in machine-to-machine format to the metadata records in IMIS
        - Add same PIDs to the metadata fields that are separately provided in the MDA
        - Review of other "missing" metadata (e.g. see Re-usability)
        - Ensure consistency between the different machine-to-machine metadata formats we provide
        - Review of the data that is provided to harvesters (e.g. google datasets) to increase direct findability
        - Better APIs for searching IMIS catalogue to be created
    - Technologies: update from JSON to JSON-LD format for metadata records
        - Consider what modifications the EML format currently also used requires to become fully acceptable at machine-to-machine level and add more provenance fields.
        - Drupal-PHP for search APIs, potentially R, python
    - Planning in time: Quarter (Q)1-Q2 2020; search APIs Q2-3 2020

Accessibility

- **Standardisation of discovery metadata not uniform over entire catalogue (IMIS) or in MDA**
- **MDA and IMIS not registered**
- **No explicit policy documents**
    - Approach: As both data providers/repositories (MDA, IMIS) and service providers (EurOBIS, Lifewatch), VLIZ work will be done towards better OAI-PMH compatibility.
        - Overview of discovery metadata fields provided in IMIS and MDA: standardisation, and improvements on (new) metadata identified by EOSC projects
        - Consider the metadata schema used (JSON-LD/EML/...: see Findability)
        - Register MDA and IMIS with e.g. Fairsharing and update entry on Re3data
        - Provide the MDA/IMIS/EurOBIS policy document (including information about PID persistence and uniqueness [Findability])
    - Technologies: No special technologies required
    - Planning in time: Q1-3 2020

Interoperability

- **No registration of metadata schema in common registry**
- **No online explanation of metadata**
- **Not all data are in standard formats**
    - Approach:
        - Register with Fairsharing etc
        - Metadata are provided in XML, JSON, EML. However:
            - EML is apparently not fully interoperable: investigate what is necessary to make it so
            - JSON-LD appears to be a more accepted format than JSON: investigate using it
            - Add link to metadata format used to the metadata themselves (in OAI-PMH compliant manner)
        - Data in MDA are not (and never will be) required to be in standard formats unless they are made public/partially public and linked to an IMIS metadata record. For these:
            - Increased interoperability for future datasets will be tackled by encouraging better standardisation of submitted data (offering the standard formats of EMODnet, EurOBIS, etc), encouraging DOI/data publishing
            - For existing datasets, the process of improving their standards is and on-going job done per project, however there is no aim to reach 100% standardisation
            - Use less free-form vocabulary in the metadata via a better use of picklists from standard dictionaries (Worms, Marine regions, UKRI/ NOC, etc)
    - Technologies: Improvements will be made within existing APIs and interfaces (Drupal, PHP, Darwin Core Archive formats)
    - Planning in time: Putting data interoperability systems in place: by Q4 2020; Implementing metadata changes Q3-4 2020

Re-usability

- **Limited provenance information**
- **No DMP for MDA**
- **Re-use metadata not always consistent within metadata records and between IMIS and MDA, and with the EurOBIS and LifeWatch systems**
    - Approach:
        - Add new provenance fields to existing metadata schemas
        - A DMP for the MDA itself is not required, however we are developing DMP templates to be adopted by all data systems and projects that VLIZ touch (LifeWatch included),
        - A trawl through the IMIS catalogue to identify inconsistent and missing metadata on access and licences will be done; updates will be done with cooperation of the data owners
    - Technologies: No special technology necessary here, updates will be done within existing systems
    - Planning in time:
        - DMP: Q1 2020; provenance to be improved Q2 2020;
        - Access/licence metadata investigation to begin Q2 2020 but will take longer to enact.

3.3 EMSO

3.3.1 Overview

EMSO ERIC's activities in ENVRI-FAIR WP9 are coordinated by EMSO ERIC Central Management Office (CMO); however, the EMSO ERIC data service group, which includes participants from all EMSO ERIC regional facilities, will participate in the implementation plan as described below.

EMSO ERIC's implementation plan is inspired by software development methodologies; however, the tasks associated with the plan do not only include software development activities but also data- and systems-related activities. Traditional paradigms such as waterfall and the V model are not suitable for this plan because the systems need to be operated during the implementation and, otherwise, achieving the targeted milestones on time would be on risk. SCRUM-based Agile methodologies are not appropriate due to the nature of the project e.g., the resources are distributed across multiple heterogeneous institutions with different standards and practices.

This implementation plan has adopted a variant of the spiral model (Boehm, 1988). This risk-driven methodology allows adapting the processes during the course of the project; however, it is anticipated that the steps described in Figure 3 will be conducted for each of the refinement cycles. This also represents a variant of the traditional iterative and incremental development methodologies. However, other models may need to be adopted based on continuous risk analysis.

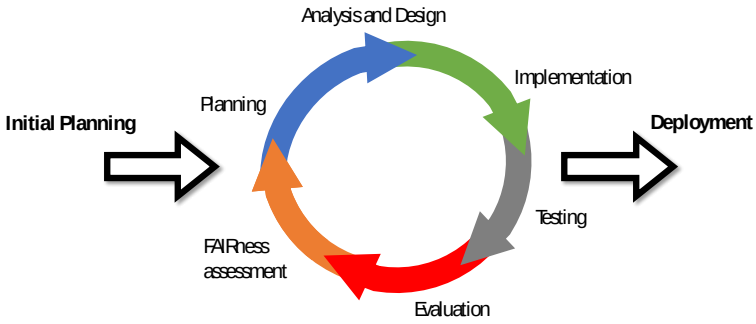

Figure 3: Variant of the spiral model (Boehm, 1988[1]) adopted in the implementation plan.

The "FAIRness level" assessment shown in Figure 3 will be conducted using different mechanisms, including:

- A self-assessment (self-evaluation) with respect to the requirements formulated in the summary of current gaps for the adoption of FAIR principles in EMSO ERIC.
- An evaluation tool developed by GO FAIR (https://w3id.org/AmIFAIR) which evaluates resources FAIRness against collections of maturity indicators tests. For example, in May 2019 the EMSO ERIC's interim solution scores 5 succeeded tests out of 22, which represents ~22%.
- Feedback from users, when possible.
- External evaluation (e.g., other research infrastructures), when possible.

[1] Boehm, B., 1988: A Spiral Model of Software Development and Enhancement. IEEE Computer, 21 (5), 62-72.

3.3.2 Gaps

EMSO has considered the following gaps as raised in report D9.1.

Findability

- Data discovery mechanisms are not harmonized across EMSO ERIC regional facilities. A current interim solution provides data discovery only via HTTP protocol, not via a metadata catalogue or standard repository solutions.
- PID registration (e.g., via DOI) is not implemented for all EMSO ERIC data/metadata.

Accessibility

- EMSO ERIC currently uses different vocabularies as part of its regional facilities data workflows as opposed to common vocabularies.
- No (consolidated) metadata harvesting is currently available in EMSO ERIC.
- Central search on metadata and data is not currently available in EMSO ERIC. Searching EMSO ERIC data using search engines does not return optimal results.
- No document describing EMSO ERIC data access protocol(s) is currently available online.
- Lack of a single point of access to EMSO data/metadata (e.g., API). EMSO ERIC already has in place web services but they are still heterogeneous. Data access is mainly via HTTP-like protocols.
- EMSO ERIC does not have a centralized user database at this point.

Interoperability

- Data and metadata are not homogenous across regional facilities. The major gap is the integration of the different regional facilities in a way that data is findable and accessible in a more uniform manner. Harmonization needs completion.
- Integration efforts are required to improve interoperability, including more mature standardization, better semantics and standard metadata.

Re-usability

- A (consolidated) EMSO ERIC repository is neither registered nor certified.
- Provenance information is partially included.
- Documents describing the persistence policy for data/metadata identifiers and metadata file format are not currently available for the entire ERIC.
- EMSO ERIC does not have an integrated data curation process at this time.

3.3.3 Implementation actions

Six refinement iterations are detailed in the execution plan schedule below; however, key ENVRI-FAIR WP9 milestones and deliverables are in project months M15, M18, M27 and M36. The schedules of each of the iterations are described below along with a short description of the goals targeted for each iteration, deliverables and tasks (see summary of current gaps). These activities are planned as part of the ENVRI-FAIR project.

The targeted FAIRness level refers to the expected percentage of successful tests using the FAIR evaluation services used when creating D9.1, and further upgraded by WP5 in cooperation with GO FAIR.

Iteration: 1 (FAIRness level > 35%)	Project Month: M15
Goals/activities	• Deployment of the harmonization subsystem based on OceanSites. The first version is expected to be partially automated. • Deployment to a baseline (refactored) machine-to-machine (RESTful) API. • Deployment of mechanisms for file-based discovery and access. • Discussion with RIs about API endpoints to facilitate interoperability across RIs in the marine subdomain for the technical specification of RI services and interfaces. • Deployment of a (refactored) metadata catalog. Although other standard repository solutions will be explored, this represents a baseline for operations. • Engagement with a provider of persistent identifiers. • Engagement with EOSC-Hub for EMSO ERIC integration in EOSC.

Iteration: 2 (FAIRness level > 40%)	Project Month: M18
Goals/activities	• Refinement of the harmonization subsystem (version 2) with a focus on integration with other components. • Refinement of the RESTful API according to the technical specification of RI services and interfaces. It may include the inclusion of common endpoints in the marine subdomain. • Incorporation of capabilities for PID management in the harmonization subsystem processes. • Implementation of mechanisms for file-based data integrity. • Refinement of the metadata catalog including both core meta-data parameters and agreed additional parameters based on an assessment of the metadata catalog against FAIR principles (emphasis on metadata enrichment). • Integration of federated identity management capabilities. • Initial scalability tests for EMSO-ERIC's integration in EOSC.

Iteration: 3 (FAIRness level > 50%)	Project Month: M24
Goals/activities	• Refinement of the harmonization subsystem (version 3) with a focus on improving the automation of processes. • Refinement of the RESTful API according to community feedback. • Refinement of the metadata catalog with a focus on search functionalities. • Deployment of software tools with data discovery capabilities that are widely used in the marine domain (e.g., ERDDAP, THREDDS) with selected data for testing. • Exploration of EOSC-hub monitoring and accounting services integration. • Registration of EMSO-ERIC services in the EOSC Portal.

Iteration: 4 (FAIRness level > 60%)	Project Month: M27
Goals/activities	• Refinement of the harmonization subsystem and other software components as needed. • Transition of software tools with data discovery capabilities that are widely used in the marine domain (e.g., ERDDAP, THREDDS) to production with initial datasets. • Evaluation of metadata for discovering EMSO ERIC data using search engines. • Registration of the EMSO ERIC in research data repositories (e.g., re3data.org). • Investigation of appropriate granularity and versioning of dynamic

	datasets according to best practices. • Development of a user database using standard tools and anonymization tools for sensitive data, when possible.

Iteration: 5 (FAIRness level > 70%)	**Project Month: M30**
Goals/activities	• Refinement of the harmonization subsystem and other software components as needed. • Configuration refinement of tools that are widely used in the marine domain (e.g., ERDDAP, THREDDS). • Elaboration of documentation related to persistence policy for data/metadata, metadata file format (e.g., URL to a vocabulary) and EMSO ERIC data access protocol(s) in the appropriate EMSO ERIC web site(s). • Investigation of widely-used and open-source digital repository software solutions (e.g., Samvera). • Engagement with the user community for feedback on the delivered interfaces, tools and services.

Iteration: 6 (FAIRness level > 80%)	**Project Month: M36**
Goals/activities	• Refinement of the harmonization subsystem and other software components (e.g., based on user feedback and/or external evaluation). • Integration of tools that are widely used in the marine domain (e.g., ERDDAP, THREDDS) with the harmonization subsystem. • Investigation of mechanisms for adding provenance information as part of the harmonization subsystem processes. • Investigation of mechanisms for certification of repositories (e.g., CoreTrustSeal requirements). • Elaboration of project documentation (e.g., deliverables).

3.4 Euro-Argo

3.4.1 Overview

Euro-Argo concluded from the FAIRness analysis that the system is well underway to be FAIR, but more FAIR to people than to machines.

Findability

Argo data DOI is well findable: description and direct link to the whole Argo dataset (with its documentation and a history of monthly snapshots).
However, only an individual person can read the documentation and identify specific observations such as a float vertical profile for a parameter.
The metadata are well detailed, but only described in the user manual not in a machine-to-machine friendly "Vocabulary server".
Only part of the metadata is searchable either on Global Portals (GDACs) or at JCOMMOPS.

Accessibility

All data are available at GDACs identified in a unique DOI.
All data are open and free so no authentication and authorization through "http", "ftp", Thredds or ERDDAP protocols.
FTP is not a FAIR protocol, one needs to know the data/metadata organization.
Thredds and ERDDAP are better but not yet robust enough for significant queries on 20 years of data.

Interoperability

NetCDF is self descriptive as long as metadata are described in Vocabulary services (which is not available).

Re-usability

Difficult to understand the processing without the QC User manual. There is a need to enhance metadata on error estimation or level of QC.

Taking this as starting point, the FAIRness improvement of Euro-Argo will focus on the back office layer, as illustrated in the next diagram.

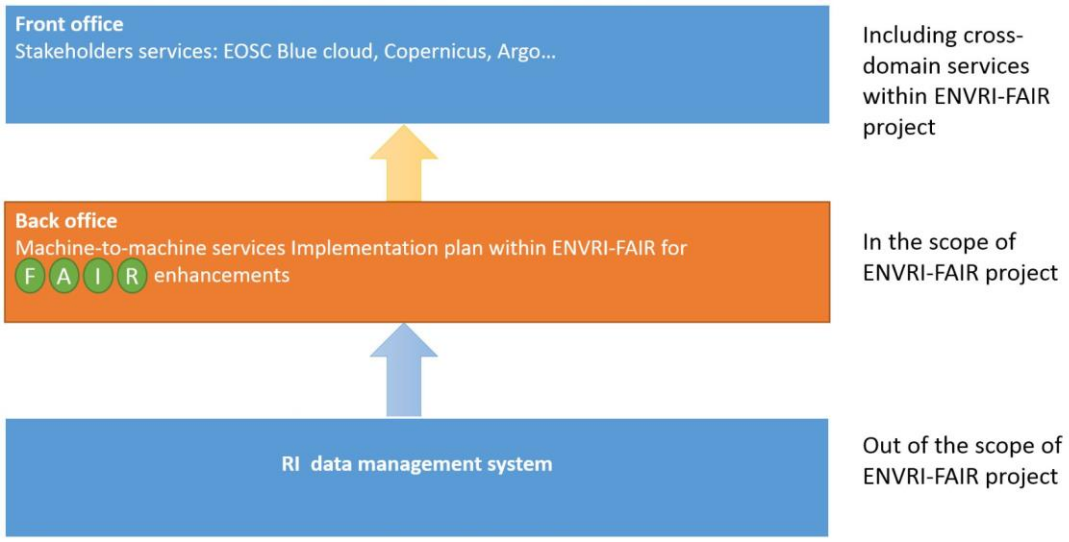

Figure 4: Illustration of focus for FAIRness actions.

3.4.2 Gaps

Euro-Argo has considered the following priority weaknesses in the gaps raised in report D9.1.

Findability

- **Weakness 1:** the global Argo dataset is well findable. However, a rich and efficient search service at data level is missing.

Accessibility

- **Weakness 2:** on top of the collection of NetCDF files, API cloud services should be developed/enhanced and implemented for data discovery, visualization, download, subscription with rich sub-setting for:
  - Individual scientist end users, with direct access to the original Argo NetCDF files, through a web GUI (Graphic User Interface)
  - Virtual Research Environment with cloud data and metadata (API queries instead of queries on transferred NetCDF files)
- **Weakness 3:** improvements on subsetting services with OGC WMS and SOS services.
- Introduce semantic web capabilities through Linked data/SparQL endpoint development.

Interoperability

- **Weakness 4:** Argo vocabulary tables are not yet implemented in the SeaDataNet vocabulary service, and not yet linked with other vocabularies.
  Euro-Argo will implement a vocabulary server for Argo metadata to solve this. The provenance of data in the Argo Data System is underpinned by rich metadata which is standardised across the data system using vocabularies currently documented in manuals and associated spreadsheets.
  The accuracy, controlled evolution and semantic value of this metadata will be further enhanced by migrating to a controlled vocabulary management environment and server compliant with W3C standards.
  With the vocabulary server, a proper and registered Argo schema will be published.
- **In the framework of Task 9.8 (EOV global product) :** Euro-Argo will implement a data processing HUB of public and shared codes for tools such as data processing chains (Argo floats decoders), file format checkers, decimation or standardization services.

Re-usability

- Provenance info is now text only, should be coded and registered in the SeaDataNet Vocabulary service (see **Weakness 4** above).

3.4.3 Implementation actions

Focused on "Back office" machine to machine services, Euro-Argo has selected seven priority actions on identified weaknesses to improve FAIRness within the ENVRI-FAIR project. These weaknesses are addressed in the orange boxes of the diagram below.

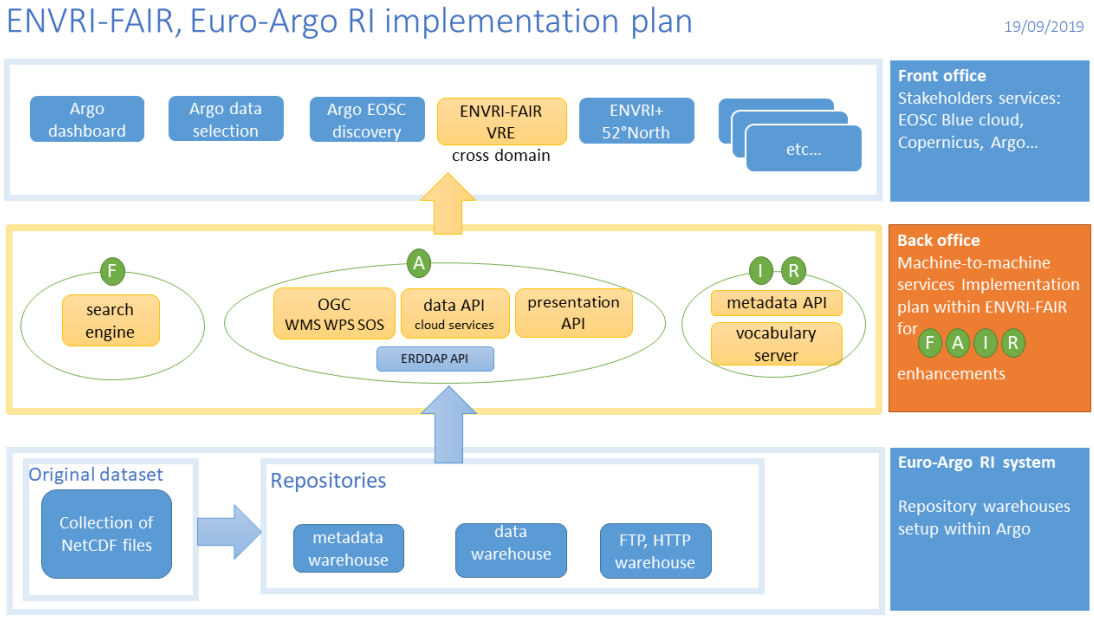

Figure 5: Euro-Argo's areas for implementation of F A I and R.

Findability

- **Weakness 1: findability - no open query on Argo data**
  - The suggested approach is to implement a search engine service
  - Possible technologies: OpenSearch on top of an Elasticsearch metadata repository
  - Planning in time: Proof of concept (POC) in Q2 2020, demo version Q1 2021
  - Involved partners: IFREMER, WP5-7 support (MARIS, UvA)

Accessibility

- **Weakness 2.1: accessibility – no machine to machine metadata access**
  - The suggested approach is to implement a metadata API
  - Possible technologies: Elasticsearch or SOL-R on metadata repository
  - Planning in time: POC in Q2 2020, demo version Q1 2021
  - Involved partners: IFREMER, WP5-7 support (MARIS, UvA)

- **Weakness 2.2: accessibility – no machine to machine data access repository**
  - The suggested approach is to implement a data API
  - Possible technologies: Cassandra, Parquet or HBASE on data repository
  - Planning in time: POC in Q2 2020, demo version Q1 2021
  - Involved partners: IFREMER, WP5-7 support (MARIS, UvA)

- **Weakness 2.3: accessibility – no machine to machine description of the above data and metadata API services**
  - The suggested approach is to implement an API description service to facilitate the use of Argo APIs
  - Possible technologies: swagger
  - Planning in time: POC in 2020 Q2, demo version 2021 Q1
  - Involved partners: IFREMER

- **Weakness 3: accessibility – not accessible through OGC protocols**
  - The suggested approach is to implement the major OGC services on top of data and metadata APIs: WMS for map services, SOS v3 for data queries, WPS to activate filtering and subsetting of the data processing HUB (ENVRI-FAIR VRE)*
  - Possible technologies: 52North or Geomatys (geoserver) solutions
  - Planning in time: POC in Q2 2020, demo version Q1 2021
  - Involved partners: IFREMER

Interoperability/re-usability

- **Weakness 4: interoperability/re-useability – not interoperable with other RIs and scientific domains**
  - The suggested approach is to implement a vocabulary server for metadata
  - Possible technologies: skos queries on a vocabulary server
  - Planning in time: POC in Q2 2020, demo version Q1 2021
  - Involved partners: UKRI/ NOC, IFREMER

(*) Data processing demonstration in the framework of Task 9.8 – EOV global product

  - The demonstration will implement the data processing on Cassandra repository
  - Possible technologies: Scala, python, Spark, SparQL query languages on top of Casandra or parquet repositories
  - Planning in time: POC in Q1 2021, demo version Q1 2022
  - Involved partners: IFREMER, UiB, OGS, UKRI/ NOC, MARIS, VLIZ, INGV

3.5 SeaDataNet

3.5.1 Overview

Following the results of the questionnaires and first FAIRness analysis, the focus for SeaDataNet will be on its CDI data access service, and the SeaDataNet data product catalogue. And similar to Euro-Argo most attention will go to the service layers for machine to machine access, while the dataflow from the sources and the data repositories will remain out-of-scope (being internal to the SeaDataNet community). This is illustrated in the diagram below.

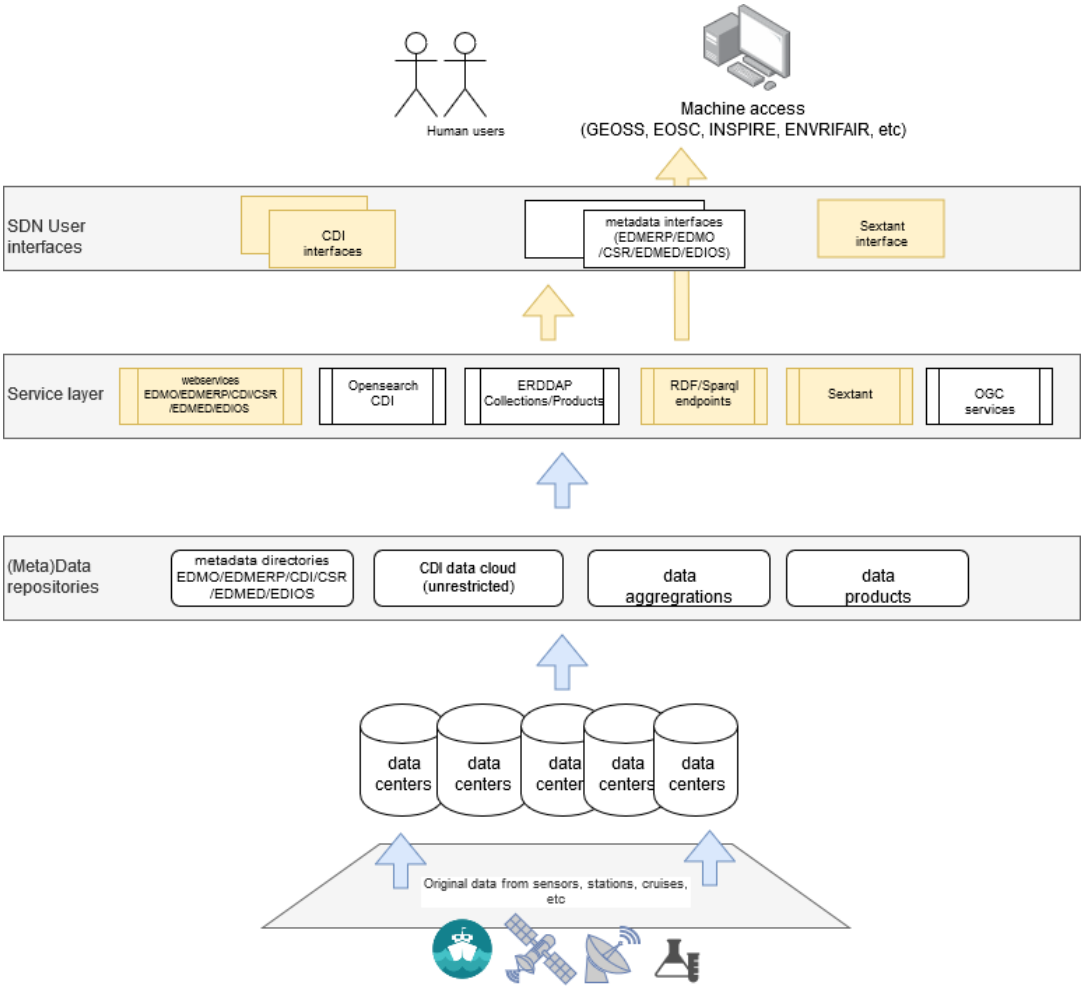

Figure 6: Schematic dataflow in SeaDataNet RI with focus areas for FAIRness.

3.5.2 Gaps

SeaDataNet RI partners have considered the following gaps identified in report D9.1.

Findability

- none

Accessibility

- Sextant and CDI: search on data via ERDDAP (planned for SeaDataCloud => out of scope)
- CDI: export formats – Json and RDF (planned for SeaDataCloud => out of scope)
- CDI: metadata exchange/access - Extension with SPARQL endpoint (planned for SeaDataCloud in synergy with ENVRI-FAIR)
- CDI: API development for improved machine access
- Sextant: Registration of repository in official registry
- Sextant and CDI: Explicit persistency policy in metadata
- Sextant: Missing access to dataplots via WPS

Interoperability

- CDI: Storage format of metadata: Extension with RDF - SPARQL
- Sextant: Not all categories in metadata marked up with vocabularies, some still in text.

Re-usability

- CDI: Categories mostly supported by controlled vocabularies, but a few with free text (especially concerning provenance and quality)
- Sextant and CDI: Provenance info - Metadata enrichment planned for structured provenance metadata, using linked data principles and SeaDataNet directories and vocabularies. Machine readable and interpretable information needed e.g. for quality info, processing info. This info is known but not captured and provided in structured metadata.

3.5.3 Implementation actions

Taking into account ongoing work in the SeaDataCloud project on several of the gaps, the overview below indicates the actions undertaken to be undertaken as part of ENVRI-FAIR. From the SeaDataNet community there is involvement in ENVRI-FAIR from IFREMER, MARIS and UKRI/ NOC. It is important to note that these partners have the expertise and experience to work on the solutions indicated and therefore have been assigned to these actions.

Accessibility

- **CDI accessibility: API development for improved machine access**
  - Suggested approach: develop a restful API to improve the ordering and download of CDI (unrestricted) data by machine users.
  - Possible technologies: Restful API in combination with OpenSearch
  - Planning in time: POC in 2020 Q2, demo version 2021 Q1
  - Involved partners: MARIS, IFREMER

- **Sextant accessibility: Registration of repository in official registry**
  - Suggested approach: Identify important registries, then register.
  - Possible technologies: -
  - Planning in time: Registration 2020 Q2
  - Involved partners: IFREMER

- **Sextant and CDI: Explicit persistency policy in metadata**
  - Suggested approach: Draft persistency policy, then add in CDI and Sextant ISO19139 datamodel.
  - Possible technologies: -
  - Planning in time: Drafted and concept of integration 2020 Q2, implemented 2021 Q1
  - Involved partners: IFREMER, MARIS

- **Sextant: Missing access to dataplots via WPS**
  - Suggested approach: Develop WPS to allow dataplot access, accessible from Sextant metadata
  - Possible technologies: WPS
  - Planning in time: POC in 2020 Q2, demo version 2021 Q1
  - Involved partners: IFREMER

Interoperability

- **Sextant: Not all categories in metadata marked up with vocabularies, some still in text (especially concerning provenance and quality).**
  - Suggested approach: upgrade Sextant metadata to incorporate more vocabularies. Collect provenance information from data product. After this metadata records for products and aggregations can be updated
  - Possible technologies: ISO19139, NERC vocabulary service (NVS)
  - Planning in time: POC in 2020 Q2, demo version 2021 Q1
  - Involved partners: IFREMER, UKRI/ NOC

Re-usability (focus on provenance and metadata enrichment)

- **CDI: Expand current CDI metadata with structured provenance metadata, using linked data principles and SeaDataNet directories and vocabularies (linked to cruises, projects, organisations, data sets, etc).**
  - Apart from this, machine readable and interpretable information is needed e.g. for quality info, processing info. This info is known, but not captured and provided in structured metadata. Upgrade CDI metadata eventually to incorporate more vocabularies, regarding provenance and quality.
  - Possible technologies: RDF, SPARQL, ISO19139, NVS
  - Planning in time: POC in 2020 Q2, demo version 2021 Q1
  - Involved partners: MARIS, UKRI/ NOC

- **Sextant: Collect and expose structured provenance metadata for data products and aggregations in Sextant.**
  - This information is known but not captured and provided in structured metadata yet (usually only text/pdf format). Work will be done to extract elements and store it as machine readable and interpretable information e.g. for quality info, processing steps, validation, software versions.
  - Possible technologies: ISO19139, NVS
  - Planning in time: POC in 2020 Q2, demo version 2021 Q1
  - Involved partners: IFREMER, MARIS, UKRI/ NOC

4. Conclusion

This deliverable 9.2 gives a concise overview of the implementation plans of the RIs in the marine subdomain to improve their FAIRness as follow-up to the earlier strengths and weaknesses analyses as reported in deliverable 9.1. Moreover, it points out the identified gaps for which activities will be undertaken as part of the ENVRI-FAIR project.
These RI draft implementation plans have already been discussed in the WP5 review meeting of 29 October 2019 (together with WP7-11) and there have provided very useful input which has contributed to establishing taskforces for the following themes:

- ENVRI (VO) AAI implementation
- PID's, identification, types and registries
- Triple stores, data storage, certification
- Licenses, citation, usage tracking
- User feedback and demonstration cases in Jupyter
- ENVRI catalogue of services

The next deliverable for WP9 will be D9.3 by March 2020 in which for each of the RIs in the marine subdomain more detailed technical choices, priorities in development, and planning will be reported. These analyses to progress from D9.2 to D9.3 will be undertaken by the RIs interacting and participating in these ENVRI-FAIR WP5/WP7 taskforces.

Appendix A: Glossary

API	Application Programming Interface
B2HANDLE	EUDAT minting, storing, managing and accessing persistent identifiers
CAS	Central Authentication Service
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
CSR	Cruise Summary Report
CSW	Catalogue Service for the Web
DMP	1) Data Management Plan 2) Data Management Platform (WP9)
DOI	Digital Object Identifier
DSA	Data Seal of Approval
ECV	Essential Climate Variable
EDIOS	European Directory of ocean Observing Systems
EDMED	European Directory of Marine Environmental Datasets (SeaDataNet)
EDMO	European Directory of Marine Organisations
EDMERP	European Directory of Marine Environmental Research Projects
EMODNET	European Marine Observation and Data Network
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
ENVRIplus	An environmental RI cluster H2020 project
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
FAIR	Findable Accessible Interoperable Reusable
GBIF	Global Biodiversity Information Facility
GCMD	Global Change Master Directory
GDAC	Global Data Assembly Center
GEMET	GEneral Multilingual Environmental Thesaurus
GEO	Group on Earth Observation (System of Systems)
GEOSS	Global Earth Observation System of Systems
GOFAIR	An international programme on FAIR implementation
GUI	Graphical User Interface
ICOS	Integrated Carbon Observation System
ICT	Information and Communications Technology
IMIS	Integrated Marine Information System
INSPIRE	Infrastructure for Spatial Information in the European Community
iRODS	Open Source Data Management Software
JCOMM	Joint Technical Commission for Oceanography and Marine Meteorology
M	Month

Marine-ID	Registration and authentication services for marine data services
MDA	Marine Data Archive
NetAPP	Hybrid cloud service
NetCDF	Network Common Data Format
NVS	NERC Vocabulary Services
NOAA	US National Oceanic and Atmospheric Administration
OAUTH	Open Authorization (standard)
OAI-PMH	Open Archives Initiative Protocol for Metadata Harvesting
OBIS	Ocean Biogeographic Information System
OGC	Open Geospatial Consortium
OpenDAP	Open-source Project for a Network Data Access Protocol
ORCID	Open Researcher and Contributor ID
OWL	Web Ontology Language
PID	Persistent Identifiers
POC	Proof of Concept
PROV-O	Web Ontology Language encoding of the PROV Data Mode
Q	Quarter
QA/QC	Quality Assurance/Quality Control
RDF	Resource Description Framework
RI	Research Infrastructure
RSS	Really Simple Syndication
SAML	Security Assertion Markup Language
SEADATANET	SeaDataNet pan-European infrastructure for marine data management
SME	Small or medium Enterprise
SparQL	SparQL Protocol and RDF Query Language
SWOT	Analysis on Strengths, Weaknesses, Opportunities and Threats
VRE	Virtual Research Environment
WoRMS	World Registry of Marine Species
WPS	Web Processing Services