D7.5
Report on the co-development and technical workshop, V1

Work Package	WP7
Lead partner	UvA
Status	Final
Deliverable type	Report
Dissemination level	Public
Due date	30-06-2021
Submission date	29-06-2021

Deliverable abstract

The overarching goal of ENVRI-FAIR is for all participating RIs to improve their FAIRness and prepare the connection of their data repositories and services to the European Open Science Cloud (EOSC). With the development of FAIR implementations from the participating RIs and integrated services among the environmental subdomains, these data and services will be brought together at a higher level (for the entire cluster), providing more efficient services for researchers and policymakers.

This deliverable reports the efforts WP7 contributed to the development of common FAIRness goals.

The objective of this task is to provide co-development support driven by an implementation plan that identified by each subdomain, joint use cases among RIs or subdomains, and other needs identified during the 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	Zhiming Zhao	UvA	01-05-2021
Contributing Authors	Siamak Farshidi Markus Stocker Barbara Magagna Markus Fiebig Keith Jeffery Peter Thijsse Christian Pichot Nicola Fiore	UvA TIB EAA NILU NERC/EPOS MARIS/SeaDataNet INRA/ANAEE LifeWatch ERIC	
Reviewer(s)	Maggie Hellström Alex Vermeulen Angeliki Adamaki	ICOS/LU ICOS/LU ICOS/LU	21-06-2021
Approver	Andreas Petzold	FZJ	29-06-2021

DELIVERY LOG

Issue	Date	Comment	Author
V 0.1	12-04-2021	First Draft	Zhiming Zhao

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

D7.5 - Report on the co-development and technical workshop, V1

1 Introduction

Task 7.3 provides effective consultancy and support for the FAIR development and technology transfer activities within individual RIs, adopting or customizing existing technologies and investigating reusable solutions for common problems that emerge at the cluster and individual RI levels.

The task performs the following actions:

- Provides consultation and guidance for adopting services inherited from earlier projects, e.g., ENVRIplus as well as those provided/promoted by relevant mature projects, such as EOSC Pilot and EOSC Hub
- Conducts collaboration activities (workshops, face-to-face meetings, etc.) with RIs to put work into practice
- Supports the RIs in implementing services for enabling FAIR data: a) implementation of FAIR metadata and data, b) implementation of services for data findability and access, c) implementation of high-level services and d) implementation of FAIR data interoperability and re-use.

This task covers joint FAIR adoption, customization, and recommendations. Specifically, T7.3 advises RIs on adopting services, co-develops required services with the RIs, and supports RIs in implementing those services. Essentially, T7.3 takes services in an architectural setting recommended by WP5 and the task forces, co-develops these together with the subdomains, and supports their implementation by RIs. Due to the support nature of the WP7, it is not always easy to separate activities of WP7 from WP5, the subdomains WP8-11 and the task forces, since almost all the WP7 members have multiple roles in the project.

This deliverable summarizes the methodology and activities performed by WP7 in the context of this task, reports the key output during the past two years, and plans the activities in the rest of the project.

2 Challenge and methodology

2.1 Challenges

In the ENVRI-FAIR project, we face a highly diverse landscape of the current ENVRI research infrastructures. First of all, research infrastructures in each subdomain have different maturity levels of their data management service. Some are already operational for quite some time, e.g., EuroArgo; some are still in the early phase of the development, e.g., DiSSCo and DANUBIUS. Moreover, the priority and timeline of specific data management services are usually different across RIs, since a RI has to serve not only needs from its scientific communities but also the international (global) network to which they are connected, e.g., EuroArgo is part of the Argo network. In this context, the standards, vocabularies, and practices used in RIs from different subdomains are usually different. The developments of the FAIR data and services in those RIs need to consider the e-infrastructures, like EOSC, for future deployment or operation. The rapid evolution of those external infrastructures provides rich technical choices for underlying services, but their insatiable and rapidly changing nature makes the integration processes difficult.

To effectively support FAIR data and service development, WP7 has to face challenges. FAIRness developments are initiated bottom-up, so WP7 must start by precisely capturing the common part of the implementation plan among the RIs in the various ENVRI subdomains, which is challenging. Limited development resources in WP7 makes it challenging to contribute to many development efforts in different RIs. Each RI and subdomain has a multitude of developers, each of them possibly located in different institutions, and they often have disjunct agendas for development activities. Therefore, it is time-consuming to have an effective discussion among researchers from multiple RIs.

2.2 Methodology

We have chosen the following methodology to organize the activities in WP7:

1. **Close engagement** in key activities may contribute to the FAIR development in subdomains, including FAIRness assessment, technology exploration (via WP5 Task Forces), training, and use case development.
2. **Regular meetings** with subdomains help WP7 to get the latest update from RIs.
3. **The agile approach** was used to establish a pragmatic way to tackle technical problems, e.g., in use cases or specific technologies.
4. The **ENVRI community Knowledge base** was prioritized as an instrument to document and share knowledge and practices from subdomains. It will enable user-centered information discovery from different ENVRI sources.

3 Report of the activities

3.1 Identify common activities

Common activities were identified from different aspects:

1) implementation activities planned by each subdomain (via specific deliverables),
2) individual discussion WP7 had with each subdomain on their use cases,
3) activities in different task forces,
4) training activities,
5) observations from interactions the ENVRI community has with the EOSC, RDA and other initiatives.

In this section, we will briefly summarize the common activities we identified from each aspect.

3.1.1 Common activities identified from subdomain implementation plans

Based on the implementation plans from each subdomain in ENVRI-FAIR (D8.3[1], D9.2[2], D10.2[3], and D11.1[4]), we summarize the common activities from those plans in a live table[5]. The initial version has also been checked against the common implementation plans summarized in the D5.2[6]. From this summary, we can highlight the following topics that have been highlighted by at least one subdomain:

a. *Metadata standardisation,* e.g., for standardising the metadata interface (Atmosphere), for assuring reusability for data products, like on ocean acidification (Marine), for rich metadata catalogue (Solid earth), and for enabling data discovery (Ecosystem). The metadata standardisation activity is often closely connected with the standardisation of the domain vocabularies and ontology, as we can clearly see from the implementation plan of each subdomain. Furthermore, standardisation within each subdomain is needed for the ENVRI catalogue of services (in D5.2). The conversion will be done using EPOS-DCAT-AP as an example of DCAT-AP implementation to a final CERIF[1] catalogue.
b. *Semantic search*, using the metadata and data access services as developed and exposed (RESTful API's, Sparql endpoints mainly), e.g., using RDF (Ecosystem), following metadata standards like EPOS-DCAT-AP and CERIF (Solid earth), via software like ERDDAP and THREDDS (Marine), and semantic search interface in the portal (Atmosphere). All subdomains address the issue of metadata mapping (among data, field or other types of assets), which is a key enabling technology for enabling search across catalogues.
c. *Persistent Identifiers* have been highlighted by almost all subdomains, e.g., for organizations and devices (Atmosphere), for findability and accessibility (Marine), for policies for PID

[1] https://doi.org/10.5281/zenodo.3885240
[2] https://doi.org/10.5281/zenodo.3885327
[3] https://zenodo.org/record/3925503#.XzDrDxMzZBw
[4] https://doi.org/10.5281/zenodo.3885361
[5] https://docs.google.com/spreadsheets/d/1JINwEC1gSLZmJlimSBMQg37t610kahiBvKj_ClosxWA/edit?usp=sharing
[6] https://zenodo.org/record/4061702#.X37yHJMzZQI

(Solid earth), and for improving metadata of PID (Ecosystem). PID usage throughout the data production workflow, including provenance, will be an important common task.

d. *Provenance*, e.g. consistent documentation throughout data production workflow (Atmosphere), improving the metadata model for provenance capturing (Marine), prototype provenance template (based on PROV) in computational workflow and prototyping using CERIF (Solid earth), and improving metadata for supporting provenance (Ecosystem). The provenance standards, e.g., PROV-O, have been considered by most of the RIs.

e. *Catalogue* activity has been highlighted by almost all RIs in four subdomains. A shared vision behind the recommendations from WP5 TF1 is a cluster-level catalogue using EPOS-DCAT-AP compliant metadata as the common conversion format leading to a CERIF common catalogue in the context of the emerging ENVRI-hub.

f. *AAI* (authentication and authorization infrastructure, sometimes also used as authentication, authorization, and accounting infrastructure) is crucial for enabling access to data and service assets provided by the remote infrastructure. In ENVRI, the authentication part has been supported by several RIs, e.g., using SAML and/or OIDC/OAuth2 (e.g., in ICOS), or using Marine-ID (e.g., in Marine). Some RIs offer free data services (as well as restricted) but are considering the AAI solution for future operation, e.g., SIOS and EPOS. A community-level agreement on user authentication, e.g., using ORCiD or AARC, ENVRI Virtual Organization (VO), or recommendations from EOSC, will be needed. The current recommendation from WP5 TF2 is to use OIDC/OAuth2,

3.1.2 Common activities in the use cases from subdomains

We had individual discussions with each subdomain on their use cases. We aim to better understand the use case scenarios, identify their common patterns, and plan actions for WP7 to join the effort.

3.1.2.1 Atmosphere subdomain

The FAIRness implementation plan of the ENVRI-FAIR atmospheric subdomain is based on two data FAIRness assessments, a preliminary one conducted while writing the ENVRI-FAIR project proposal, and a second more thorough one conducted during the first 6 project months of ENVRI-FAIR with the guidance of WP7 experts. The assessment resulted in a data FAIRness implementation plan focussing on the following aspects:

- Use of suitable PIDs for all entities involved in data production
- Common standardised metadata interfaces for metadata and data access
- Indexing of data resources and services in domain-specific search engines, as well as ENVRI-hub
- Common use of authentication schemes
- Recommendations for graphical user interfaces for data search and access
- Documentation of provenance throughout the whole data production process.

To illustrate the close collaboration between the atmospheric subdomain organised in WP8 and ENVRI-FAIR data management experts in WP7, two use cases are described in more detail.

Provenance in data production

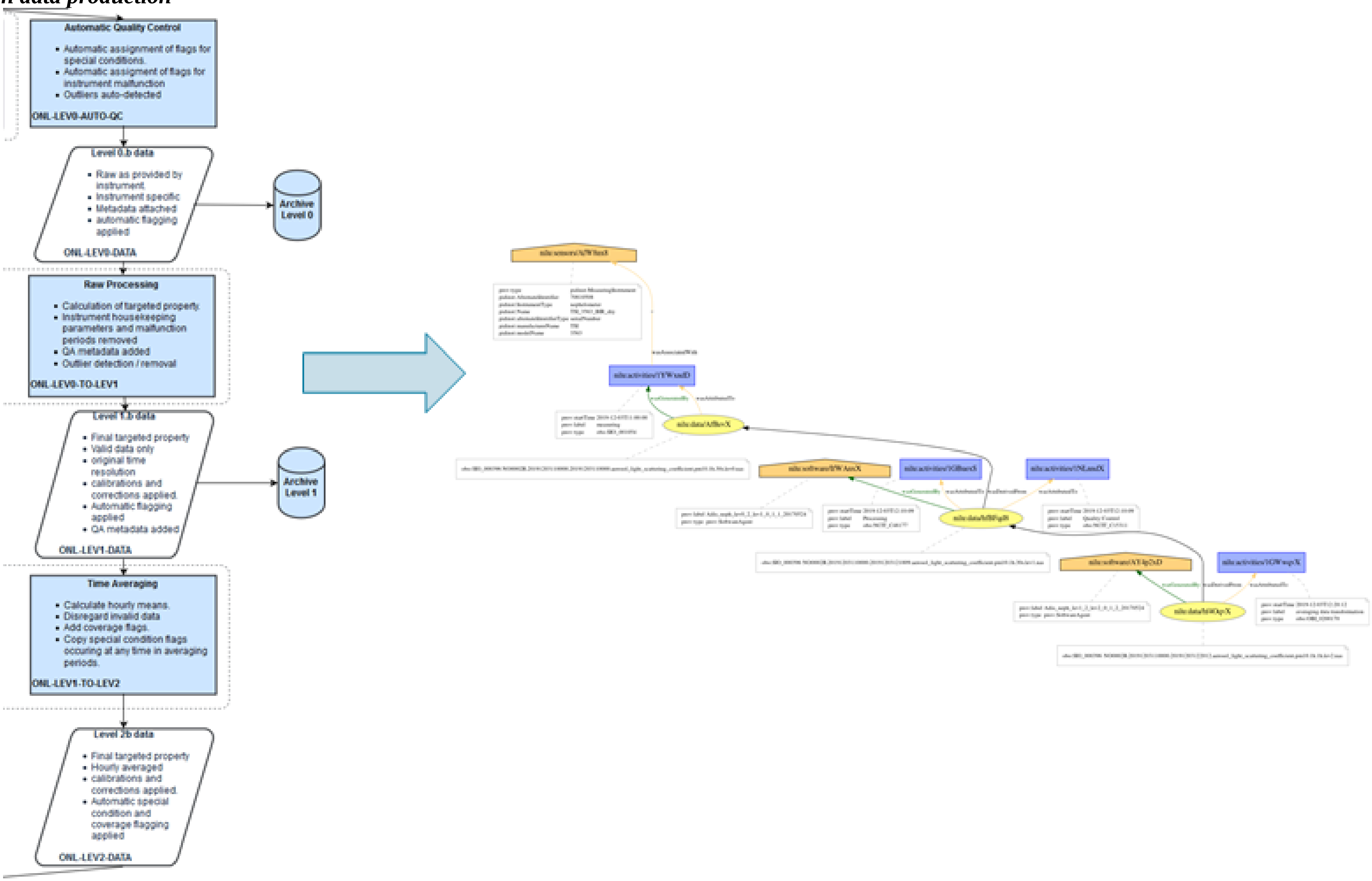

Automatic Quality Control
Automatic assignment of flags for special conditions.
Automatic assigment of flags for instrument malfunction
Outliers auto-detected
ONL-LEV0-AUTO-QC
Level 0.b data
Raw as provided by instrument
Instrument specific
Metadata attached
automatic flagging applied
ONL-LEV0-DATA
Archive Level 0
Raw Processing
Calculation of targeted property.
Instrument housekeeping parameters and malfunction periods removed
QA metadata added
Outlier detection / removal
ONL-LEV0-TO-LEV1
Level 1.b data
Final targeted property
Valid data only
original time resolution
calibrations and corrections applied.
Automatic flagging applied
QA metadata added
ONL-LEV1-DATA
Archive Level 1
Time Averaging
Calculate hourly means.
Disregard invalid data
Add coverage flags.
Copy special condition flags occuring at any time in averaging periods.
ONL-LEV1-TO-LEV2
Level 2b data
Final targeted property
Hourly averaged
calibrations and corrections applied.
Automatic special condition and coverage flagging applied
ONL-LEV2-DATA

Figure 1. Blue-print example for documenting data provenance in the atmosphere sub domain. A workflow piece, here from the ACTRIS RI (left), is used to provide an example of provenance documentation (left).

To facilitate the uptake of concepts from information sciences in the atmospheric subdomain RIs, working groups (WGs) have been set up in WP8. The WGs are staffed by members of all involved RIs as well as WP7 experts. In order to facilitate the uptake of concepts for documenting data provenance, the provenance WG developed blue-print examples by looking at RI data production workflows, and documenting execution examples by using the provenance ontology (PROV-O).

The approach proved successful to a degree that the final report of the provenance WG has now been turned into an implementation task for the atmospheric subdomain.

Subdomain reference vocabulary for observed variables

In the start of ENVRI-FAIR it became already clear that the lack of a common vocabulary is an essential hindrance for machine-actionable interoperability in the subdomain. In the course of the project the vocabulary topic has appeared in all ENVRI subdomains, leading to a project-wide Task Force within WP5. For the atmospheric subdomain, guidance by WP7 experts has led to a concept of two components:

- Enhancing the vocabulary of the Climate Forecast Convention as a reference vocabulary for the subdomain because of its exact definitions, good governance, and machine-readability.
- Taking into use the RDA I-ADOPT framework[7] for naming of observed properties in order to assemble vocabularies adapted to the detail level needed by the use case while maintaining machine-actionable relations.

3.1.2.2 Marine Subdomain

In the marine subdomain, the FAIRness analysis has led to a series of FAIRness improvements with emphasis on the Interoperability and the Reusability. As an example, an important case in the marine subdomain is the use case to support reliable studies, for which it is necessary to be able to distinguish between "premium" data and “provisional” or “routine” data:

- Premium: data acquired with the highest standard of observations; quality controlled by a scientist
- Provisional data: data available, but additional processing will improve the observations
- Routine data: typically, real-time data, without information on sensor calibration or quality control.

The provenance level of metadata makes the distinction for which level of datasets is available.

Other priority activities in the RI’s in the marine subdomain (full overview in relevant WP9 deliverables[8]) are:

- Development of machine-accessible endpoints: SPARQL endpoints and Restful API’s (among which ERDDAP)
- Metadata extensions and LinkedData developments (RDF DCAT-AP compliant) to support accessibility as well as improved provenances
- Upgraded vocabulary service content to support metadata and data models and support better mapping and machine interpretability
- Updates to metadata models for better provenance capturing.

[7] https://www.rd-alliance.org/group/interoperable-descriptions-observable-property-terminology-wg-i-adopt-wg/wiki/i-adopt

[8] https://zenodo.org/communities/envri/?page=1&size=20

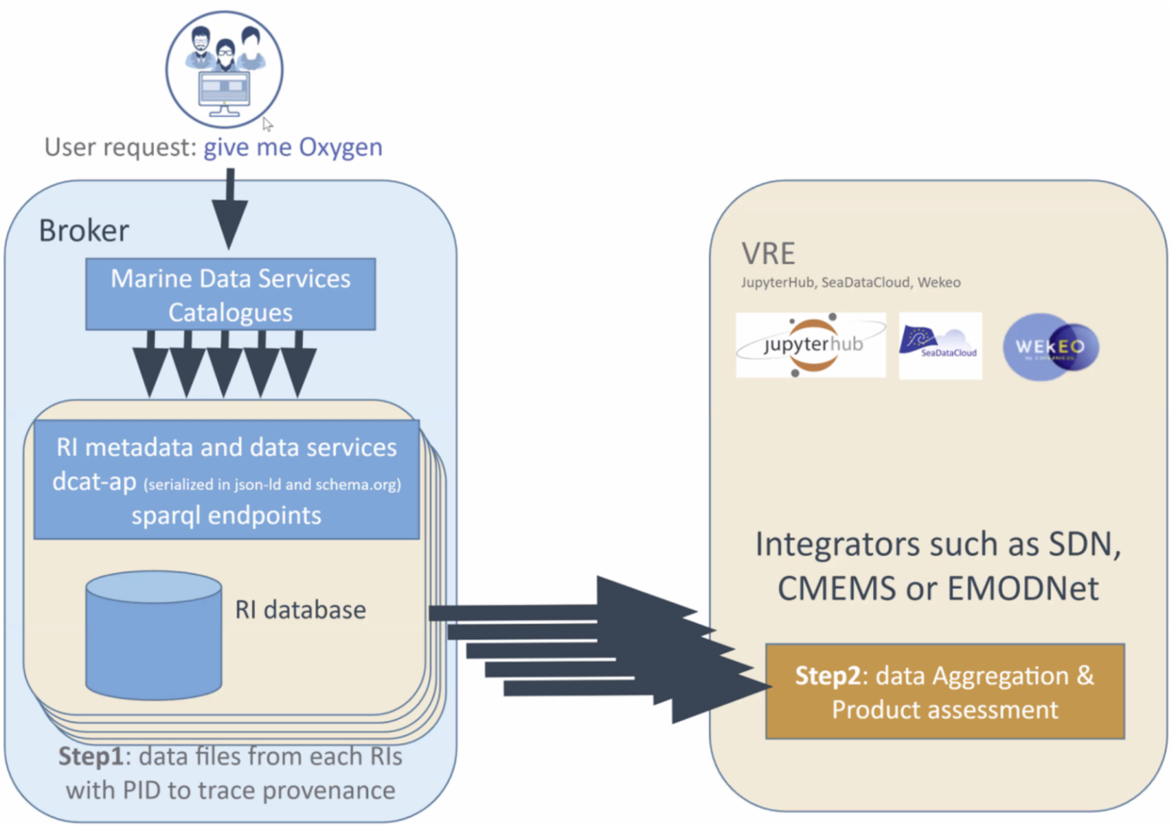
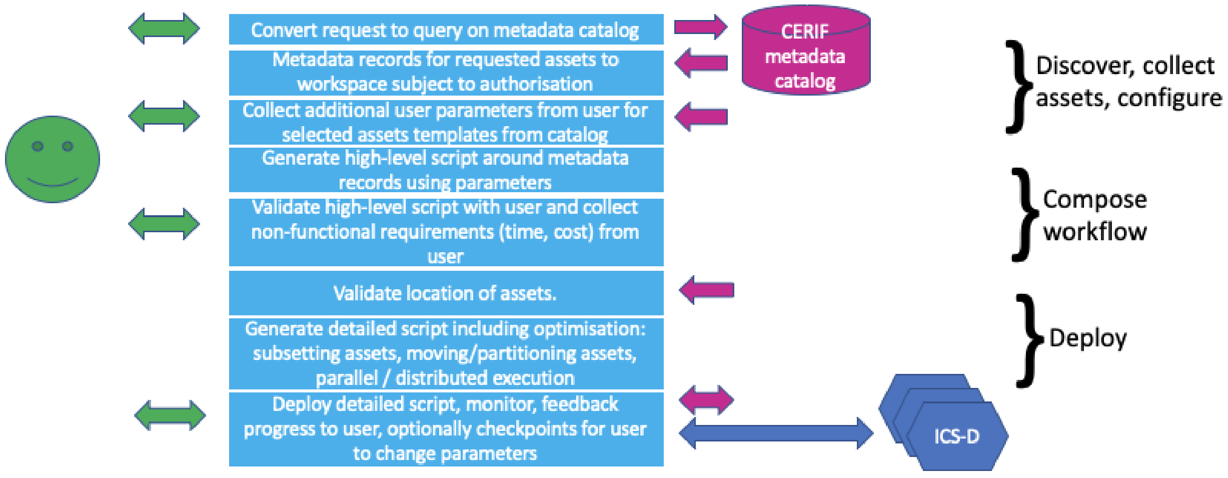

User request: give me Oxygen
Broker
Marine Data Services Catalogues
RI metadata and data services
dcat-ap (serialized in json-ld and schema.org)
sparql endpoints
RI database
Step1: data files from each RIs with PID to trace provenance
VRE
JupyterHub, SeaDataCloud, Wekeo
jupyterhub
SeaDataCloud
WEKEO
Integrators such as SDN, CMEMS or EMODNet
Step2: data Aggregation & Product assessment

Figure 2. Use case scenario in the marine sub domain.

To test and demonstrate the result of all activities above, the Marine subdomain will develop a demonstrator product querying all endpoints for the marine RI's for the same EOV (Essential Ocean Variable) data - e.g., oxygen - and be able to access and download the dataset for further use with sufficient metadata to analyse the data quality. Other uses could be in Virtual Research Environments etc.

3.1.2.3 Solid Earth subdomain

The Solid-Earth subdomain had a plan before ENVRI-FAIR, and this integrated well into the Solid-Earth subdomain implementation plan of ENVRI-FAIR. The plan is based around the principles: (a) minimal disturbance to the asset suppliers within the subdomain - in fact, the only demand is metadata supplied (by push or pull - harvesting) via supplied convertors to EPOS-DCAT-AP and thus to the EPOS central service for conversion to CERIF; (b) a central rich metadata catalogue using CERIF which is an EU Recommendation to the Member States for research information and is a superset of the 17 different metadata formats used within the asset suppliers of EPOS; (c) the catalogue ensures FAIR processing; (d) the catalogue is populated incrementally for all asset types (and other information e.g. on organizations, persons, publications…) starting with services to be rapidly EOSC-compatible; (e) the catalogue provides metadata for the assets such that workflows can be composed/orchestrated and subsequently deployed. The catalogue is accessed through an API allowing external programmatic access but also connects to the EPOS portal GUI which has powerful facilities; (f) from CERIF, the EPOS central system provides multiple exportable formats to maximize openness and interoperability. The overall EPOS concept for support of the solid-earth research activities may be presented as a workflow in Figure 3.


Convert request to query on metadata catalog
Metadata records for requested assets to workspace subject to authorisation
Collect additional user parameters from user for selected assets templates from catalog
Generate high-level script around metadata records using parameters
Validate high-level script with user and collect non-functional requirements (time, cost) from user
Validate location of assets.
Generate detailed script including optimisation: subsetting assets, moving/partitioning assets, parallel / distributed execution
Deploy detailed script, monitor, feedback progress to user, optionally checkpoints for user to change parameters
CERIF metadata catalog
ICS-D
Discover, collect assets, configure
Compose workflow
Deploy

Not yet implemented fully: work in progress

Figure 3. Use case scenario in the solid earth sub domain.

3.1.2.4 Ecosystem subdomain

The Ecosystem subdomain identified four use cases (UC) corresponding to main challenges for i) characterisation of RI experimental sites, ii) interoperability of datasets (Soil Water Content (SWC) as proof of concept) and iii) correct species scientific naming. These use cases are developed within the general context of the ENVRI-FAIR selection of DCAT as the common metadata standard for RI or subdomain catalogues to be harvested by the ENVRI-hun for the feeding of EOSC. Therefore, DCAT compliance was proposed as the 4th use case.

The ‘scientific name’ UC led by DiSSCo is the subject of a specific action relying on the survey of Ecosystem RI practices and the interaction with the main taxonomy players, in particular Catalogue of Life. Although positioned at different levels and on different objects (sites on the one hand and variables on the other) the UC 'Site documentation' (led by AnaEE and eLTER) and 'Soils Water Content' (led by AnaEE) share the common objective of providing rich and interoperable metadata.The complementarity of the expected metadata leads to link these 2 UC: the first one notably providing the general metadata for the second about the descriptive elements of the data acquisition site. It should also be noted that the ‘site documentation’ UC is linked with the TF4 task on triplestore production.

In order to properly describe sites, the DCAT standard will be extended in a similar way to what has been done by EPOS. The majority of WP11 RI using (AnaEE, LifeWatch) or being able to produce (eLTER) ISO19139 standard metadata for site description, the ISO-to-(geo)DCAT conversion has been chosen to automate the production of DCAT metadata. For the other RI, the metadata will be produced directly in DCAT format from their metadata information systems.

The SWC data set selected for the 2nd UC are managed and characterised in a heterogeneous way across RI (AnaEE, ICOS, LifeWatch, eLTER, SIOS, Danubius). Metadata descriptors were identified and concern variable, sensor, experimental context, curation or processing information and general fields about the dataset (authors, publication, licence etc.). A metadata semantic model based on the sharing or alignment of vocabularies is a prerequisite for semantic interoperability notably for the observable properties. It has been proposed that the exploitation metadata describing the SWC data set will be provided (or converted) as Ecological Metadata Language (EML) 2.2.0 records [8]. However, similarly to what is planned for ‘site documentation’, the discovery metadata will also be generated as DCAT records with additional extended metadata fields.

The development of a prototype user interface is planned as a two-step process: i) querying tools giving access to the data set and discovery metadata and ii) harvester providing harmonized data sets from different resources and exploitation metadata.

A joint action between the two use cases will aim to cross-link the two catalogues (sites and variables). WP11 will only temporarily host and exhibit the catalogues, for the purposes of developing the case studies and demonstrating their functionalities. Their future valorisation will have to be taken over by ENVRI-FAIR through a dedicated action led by WP7.

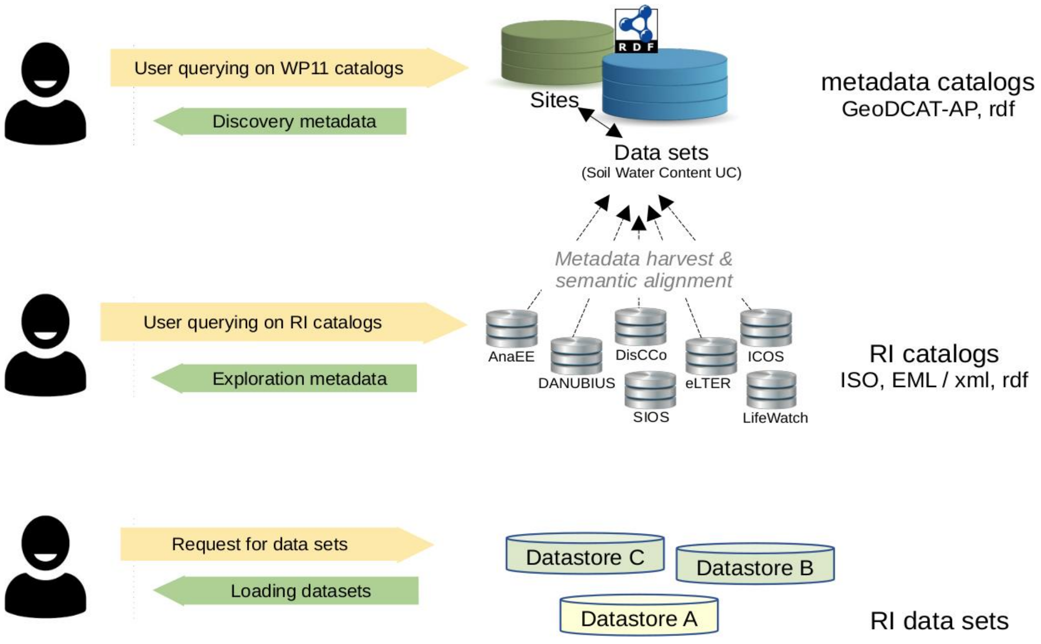


Figure 4. Use case scenario in the ecosystem subdomain.

3.1.2.5 Common activities from use cases

From the individual discussion[9] with each subdomain on their use cases, we can summarize a number of common activities and development patterns:

a. DCAT-AP has been included in all subdomains as a basis for exposing metadata on services and data in a federated ENVRI catalogue. EPOS has initialized an extension of DCAT-AP as EPOS-DCAT-AP, which can be used as a very good example for other subdomains.
b. Most of the use cases address data product workflows, e.g., for Essential Ocean Variable (in Marine), Essential Water Content (in Ecosystem), …
c. Most of the use cases contain scenarios of metadata search across multiple providers (RIs), e.g., SWC variables require input from ANAEE, ICOS, SIOS, eLTER, and DANUBIUS (in Ecosystem), Jupyter workflow for processing data from multi-sources (Solid earth), the EOV variable workflow starts from a query on a metadata catalogue (Marine),
d. Most of the use cases contain data access scenarios using PIDs, as we see from the data production workflow explained in the previous bullet (c).

From these common patterns, we can see a clear alignment between use cases and the development plans:

a. Metadata standardisation, semantic search, and catalogues will be essential for enabling the data discovery and access scenarios in use cases. In this context, we clearly see the need for metadata mapping to DCAT-AP, domain vocabulary, and the ENVRI service catalogue.
b. PID and AAI will be the basis for enabling the data access required by the use cases.
c. Workflow management is addressed by all use cases but in different priority, e.g., in the Marine subdomain, workflow management for EOV in Jupyter is not prioritized within the ENVRI-FAIR scope, while EPOS has a clear plan (see 3.1.2.3) and prototypes on its Jupyter workflow including data access, processing, remote deployment, and provenance. EPOS has developed a model for workflows consisting of the stages: discovery, contextualization, selection of assets, parameterization, composition/orchestration, deployment. The first steps are implemented. Prototypes and experiments are currently underway concerning the subsequent two steps.

[9] https://docs.google.com/presentation/d/1QC1_cWRdJ6XwoLx5g49hnRy3fpjwlD5DxQX0kvtc0fw/edit#slide=id.gcbc17d58ce_0_0

3.1.3 Common activities from the task forces

WP7 has actively participated in all task forces[10] organized by WP5. Those task forces closely work with the subdomains, investigating solutions to those common challenges, subdomain developers face during their development.

TF1 ENVRI Catalogue of Services
TF1(coordinated by the WP7 co-lead) produced a white paper [3] following a landscape survey of current catalogue practices among ENVRIs. The recommendations are (1) all ENVRI RIs will supply metadata in EPOS-DCAT-AP format; (2) the ENVRI-Hub catalogue will store the catalogue as CERIF and provide appropriate access (with conversion) to EOSC. Training and hands-on sessions have been organized to assist ENVRI RIs in generating (EPOS-)DCAT-AP from their local metadata formats.

TF2 ENVRI (VO) AAI implementation
TF2 (coordinated by the WP7 co-lead) also produced a white paper [4]. This recommended the use - within the AARC2 Blueprint architecture - of OpenID (OIDC) for authentication and OAuth2 for authorization. A landscape survey is now being conducted to identify exactly the IdPs (Identity Providers) being used and to ensure a single sign-on/federated identity. The survey will also determine access permissions required: initial work indicates that for end-users it is REL and for systems administrators the full CRUDEL (**C**reate, **R**ead, **U**pdate, **D**elete, **E**xecute, down**L**oad). Detailed work continues.

TF3 PIDs identification types and registries
TF3 will liaise with the global Persistent Identification community, including relevant Research Data Alliance groups, and with the global community of research infrastructure operators and will conduct public consultations for all deliverables in these communities as deemed appropriate by the TF. Decisions in the PID TF will be based on the principle of rough consensus to be determined by the TF chairs and confirmed on the list.

TF4 Triple stores, data storage certification and FAIR vocabularies
The aim of TF4 is to share knowledge between RIs that are experienced with triple stores, related technologies and applications, and data storage certification and RIs that are planning to use these technologies in their own applications. For this, TF4 holds monthly meetings during which issues are discussed. Additionally, TF4 has collected so-called experience reports. So far, BODC, ICOS, and ANEE have shared their experience reports on triple stores and Ifremer and BODC have shared experience reports on data storage certification. Numerous other RIs plan to adopt these technologies and have thus benefited from the shared experience reports. More recently, TF4 had an important role in the discussion related to the ENVRI-hub and, together with selected RIs, the setup of a CKAN-based demonstrator showcasing a possible implementation of the ENVRI catalogue of metadata on datasets. A second task, which started in March 2021, is related to the creation and maintenance of FAIR vocabularies. The first activities focused on the presentation of the RDA I-ADOPT Framework[7] and how this could be applied in the variable descriptions in the various subdomains.

TF5 Licences citation and usage tracking (of data and VRE)
TF5 has 3 topics of common interest, including

1) Licences for metadata and data. Targeting the "R" for Re-usability in FAIR, the TF will come up with a recommendation on which metadata items should be used to document licence (and data policy) information in metadata standards.
2) Data citation. The TF defines a scheme for data identification, that is suitable for both a) tracking of data use down to the granularity of the individual Principal investigate (PII, contributing organisation, or framework; b) giving the data user an easy way of citing a dataset or the collection of data used for a use case.
3) Data usage tracking. This task will need to collaborate with indexing agencies for data use and connect with ongoing efforts in other fora. The TF is in the process of writing down its results as a report. For the task on licence metadata, a comprehensive analysis of licence metadata representation in existing metadata schemas for scientific data was conducted. The aim is to identify an existing metadata schema that can be adapted to meet the needs of the ENVRI community with respect to machine-actionable attribution requirements.

[10] https://envri.eu/task-forces-implementing-the-envri-hub/

A team within TF5 is working on the General Data Protection Regulation (GDPR). This relates to personal data. ENVRI catalogues include personal data e.g. name of service contact or of asset supplier. Mechanisms to protect personal data relate to citation and usage tracking, but also to AAAI (TF2). The intention is to produce an appropriate privacy policy and mechanisms for ensuring user consent to the use of personal data.

TF6 ENVRI-HUB

TF6 coordinates the development and implementation of ENVRI-FAIR cross-domain demonstrator services and use cases to be exposed to EOSC via the ENVRI-hub. TF6 oversees the design of the architecture and functionalities of the ENVRI-hub, which will be driven by the applications, or use cases, and user needs, respectively. The ENVRI-hub will contain three main pillars: the ENVRI catalogue of services and datasets (TF1), the knowledge base developed by WP7, and a set of use case demonstrators. TF6 monitors the evolution of services provided at the EOSC Portal.

3.1.4 Training activities and needs

D6.1[11] presents a list of training topics that are commonly needed by the ENVRI community. The list includes topics on general FAIR-related topics, e.g., FAIR principles, metrics for evaluation, and GDPR issues related to data sharing, and on research data management, e.g., AAI, cataloguing, cloud computing, workflow management, provenance, and PID. In the training work package, various materials have been developed for the community to study those topics. The current project has also included the materials from the previous ENVRI summer schools.

3.1.5 Observations from other clusters, initiatives, EOSC

While the ENVRI cluster makes progress in ENVRI-FAIR, it is also wise to understand the approaches adopted in other clusters, in EOSC and internationally. This has been achieved through WP7 participant contacts in these areas of activity. Of particular note are discussions among science clusters on metadata - partly through the SEMAF [2] co-creation EOSC project but also in 1:1 discussion, some of which within the context of RDA and especially the Metadata Interest Group.

WP7 also joins the regular discussions with WP1-4. From those discussions, several common needs from subdomains have been identified, e.g., GDPR, data access policies, and other issues that need further support from Board of European Environmental Research Infrastructure (BEERi).

3.1.6 Summary

In this section, we have identified common development activities from different aspects: implementation plan of each subdomain, use cases, task forces, training, and other sources. At a high level, the topic of metadata standardisation, semantic search PID, cataloguing, workflow management, and AAI are addressed by most of the implementation plans. We can also see how those common activities will contribute to the data production scenarios defined in the use cases from different subdomains.

For WP7, it is important to join the discussion with subdomains on their use cases and contribute to the development and validation of those common activities.

3.2 Meetings and workshops

https://docs.google.com/spreadsheets/d/1JINwEC1gSLZmJlimSBMQg37t610kahiBvKj_ClosxWA/edit#gid=1578698276

During the second year of the project, WP7 has organized/participated in different workshops, meetings with subdomains on the common activities identified above, as shown in Table 1. Besides those activities, FAIRness assessment is one of the activities all subdomains performed in the beginning of ENVRI-FAIR and will perform several other times during the project.

[11] https://doi.org/10.5281/zenodo.3885122

Common development activities	**Meetings and workshops**	**Key contributions**
FAIRness assessment	Various FAIRness assessment workshops in 2019	Join the development of the questionnaire and assist the RIs to perform the assessment.
Metadata standardisation	Partially in TF1 and TF4 meetings	In TF1 it has been agreed that all RIs will provide metadata for the ENVRI catalogue of services in an extended DCAT (using EPOS-DCAT-AP as an example) for conversion to CERIF for the catalogue. This may take some time so, to produce something within the project timescale, work has been done in TF4. In the context of the CKAN-based demonstrator for the ENVRI Catalogue of services (to be included as part of ENVRI-hub), TF4 has coordinated the DCAT standardisation of metadata about datasets with a set of RIs. As an additional task, TF4 is coordinating the standardisation of variable descriptions (with input from the RDA WG I-ADOPT)
Semantic search	TF6 (knowledge base), and with WP8 and 11.	Join the working groups in WP8 and WP11.
Catalogue	TF1 meetings	In TF1 a white paper was produced, and it has been agreed that all RIs will provide metadata for the ENVRI catalogue of services in an extended DCAT (using EPOS-DCAT-AP as an example) for conversion to CERIF for the catalogue. WP7 will assist.
PID	TF3 meetings	Join the discussion
AAI	TF2 meetings	TF2 has produced a white paper with an outline of the existing state across the RIs and planned further steps. Agreement was reached to use OIDC and OAuth2 within the AARC blueprint architecture. TF2 is currently working on more detailed plans for RIs to evolve to OIDC and OAuth2. WP7 will assist.
Provenance	Training (webinar) and technical workshop 2021.	Provided webinars, contributed to the technical workshop, and initialized follow-up activities (readiness assessment).
Training activities for development	Summer school, and webinars (e.g., Cloud computing)	Delivered lectures

Table 1. The summary of the workshops and meetings related to common activities.

3.3 Key contributions, achievements, and lessons learned

During the past two years, WP7 contributed to a number of activities related to the development of common data services.

3.3.1 Community knowledge base

An important development effort in WP7 is to document the technical knowledge from RIs and subdomains and make them FAIR for the community. The D7.3 reports detailed knowledge base architecture, technological choices, and implementation details based on early work [7]. A search engine is provided for discovering the technical practices, online assets (documents, datasets, API, and images), and FAIR assessment reports from the ENVRI community in the community knowledge base.

3.3.2 FAIRness assessment

During the first year of the project, WP7 joined the FAIRness assessment effort in each subdomain. More specifically, WP7

a. Joint the design of the FAIRness assessment questionnaire, based on the initial GO FAIR draft
b. Re-organized the questionnaire into YAML structure to produce machine readable output
c. Supported RIs from each subdomain to conduct the assessment via a number of workshops
d. Summarized the initial output using structured documents, and presented them in the knowledge base
e. Re-designed the FAIRness assessment tool together with GO FAIR into the FAIR Implementation Profile (FIP)[5]
f. Prepared for the new iterations of the assessment
g. Performed FIP workshops at the Pre-Symposium events for the FAIR Convergence Symposium 2020.

3.3.3 Contribution to the improvement of the development plan for each subdomain

WP7 reviewed the implementation plan of each subdomain and provided comments on their plans from a cluster-level point of view. During the review, the common activities from each subdomain have been analysed and collected in an internal document[12].

3.3.4 Task Force contributions

WP7 joins the effort of all task forces, in which WP7 partners lead TF 1, 2, 4, and 5, and participates in all the others.

3.3.5 Training and knowledge transfer

WP7 participants have participated both as students and teachers in training. The ENVRI-FAIR Training Catalogue is at https://trainingcatalogue.envri.eu/ and the platform at : https://training.envri.eu/.
In particular, training has been given by WP7 participants on:

a. Conversion and uploading of metadata to the catalogue using EPOS-DCAT-AP (catalogue code 043);
b. GDPR (General Data Protection Regulation) in the context of policies and AAAI (catalogue code 045);
c. Provenance introduction (not in training catalogue)
d. Provenance using CERIF (not in training catalogue)
e. Provenance using PROV (not in training catalogue)
f. Cloud computing (Contributed to the summer school 2020)

[12] https://docs.google.com/spreadsheets/d/1JINwEC1gSLZmJlimSBMQg37t610kahiBvKj_ClosxWA/edit?usp=sharing

In addition, through the WP5 Task Forces WP7 members have participated in much ‘on the job’ education through teach-ins and discussions while working together with others, especially from the subdomains.

3.3.6 EOSC early adopter program (EAP)

WP7 coordinated the effort for an EOSC Early Adopter Program (EAP) project. In this one-year project, WP7 aimed to deploy a DevOps environment, with the necessary Cloud Infrastructures and services capacity, for testing ENVRI-FAIR developments. The project contained three case scenarios:

a. Automated Cloud execution for data workflow: demonstrate it in the VREs or ENVRIs (e.g., LifeWatch or others). It would help the ENVRI community to learn the EOSC services and build practices for the other similar use cases;
b. Continuously testing and integration for ENVRI services: get familiar with the DevOps/Agile methodologies for software development, testing, and operation;
c. Notebook-based environment for FAIR data access and processing: provide the Jupyter service to users, with examples to access data sets and models, users can perform customised experiments using the notebook services, access data, store the data, publish and share the results with the others.

During the project, we developed the following two components [6]:

a. FAIR-CELLs, a Jupyter extension to enable the interactive containerization of the Jupyter Cells.
b. Cloud-Cells, a Jupyter extension to automate the cloud services (IaaS) provisioning, and container deployment.

We applied these two components in an ecology use case called LidarCloud. Two legacy programs developed in a previous project are dockerized and executed on remote cloud infrastructures via the Jupyter environment. We run the code on the EOSC IaaS together with the VMs provided by the LifeWatch ERIC. In this way, the original code can be scaled out to process much bigger datasets than its original design.

During the project, we have reviewed the DevOps tool provided by Jelastic via the EOSC marketplace. Besides the software engineering support, we investigated the cloud automation support offered by Jelastic. The output of this study has been included in the ENVRI summer and winter schools as part of the training material.

After the end of the EAP pilot, the technical development and results will be further continued:

a. In the ENVRI-FAIR project as part of the knowledge base for supporting developers from ENVRI subdomains and RIs when developing their data management services
b. In the LifeWatch ERIC as part of the Virtual Lab solutions
c. The developers will exploit the results as part of the EGI Jupyter service, which can be visible in the future marketplace;
d. We will also actively seek the possible opportunities in future projects, like EOSC-Future to further sustain the developed solution.

4 Summary

From the report, we will summarize the plan for the next phase of the project. We will also review the possible risks and mitigation strategies.

4.1 Agenda

Based on the discussion in chapter 3, we proposed the following actions and timeline:

1. **Contribute to the ENVRI-hub development.** WP7 will continue developing the knowledge pillar in the ENVRI-hub and contribute to the development of general architecture and other two key components: ENVRI catalogue and demonstrators.
    a. *Knowledge base activity* will focus on the following aspects: i) search support for new assets, including data sets and APIs (in 2021) and images (in 2022), ii) advanced ranking and recommendation solution (in 2022), and iii) distributed knowledge management pipeline with RI engaged (in 2022).
    b. *The cataloging activity*. WP7 will support this activity in the context of TF1, 4, and 6, including i) the metadata mapping activities (from RIs to ENVRI extended DCAT-AP) and ii) harvesting pipeline to a centralized ENVRI catalogue.
2. **Contribute to the development of solutions to common problems.** WP7 will actively identify the common technical problems from the common development plan and initialize agile task forces to explore solutions. Examples include
    a. Provenance readiness assessment. As a follow-up of the provenance technical workshop (in Feb 2021), we proposed an agile task force to assess the readiness level of the provenance data from different RIs. The basic approach is, i) we interview the domain experts and prepare a list of provenance queries, ii) manually check how many queries can be answered from the data sets, metadata, logs we collected from RIs, iii) assess the level of the readiness of their provenance, and iv) provide technical recommendation for bridging their gaps.
    b. Other topics of AAI, automated workflow across distributed infrastructure and distributed data sharing will also be considered.
3. **Contribute to the validation use cases.** From the common activities we identified from the use cases, WP7 will initialize the validation support activities with the use case teams from the subdomains. Support activities being considered to include
    a. Formulating and optimizing the cloud infrastructure for demonstrating the use cases;
    b. Automating the deployment of the use case services, e.g., for discovering, accessing data, and for processing them in the distributed workflows.
4. **Contribute to the training and knowledge transfer.** WP7 will join the training effort and contribute to the training material related to technical development.
5. **Contribute to the task forces.** Within those TFs, the activities related to the use cases, common development support, and ENVRI-Hub will be specifically prioritized in WP7.

4.2 Risks and strategies

The development of FAIR data management services is time-consuming, and RIs in different subdomains often have their development plan and roadmap. Given the limited capacity of the WP7, WP7 needs to prioritize the urgent actions required by the subdomain. As a result, WP7 has to face the following possible risks:

1. **Lost focus on the urgent requirements** from the subdomains can make WP7 less connected with the subdomains. WP7 has to closely follow the technical development in subdomains, track their plan, and understand the latest requirements (members of WP are also members of the subdomain WPs).
2. **Investing too much time on less critical technical problems** can create a gap between WP7 and the technical needs from the subdomains. WP7 has to continuously review the

development status from the subdomains and prioritize the technical problems to which WP7 can contribute, in the context of the use cases and the ENVRI-hub. WP7 studies the subdomain plans for FAIRness and convergence to realise the ENVRI-hub.

3. **Lost awareness of the technical development status of RIs**. Regular technical meetings with different RIs are essential.

4.3 Timeline, and expected output

The activities for the rest of the project will be organized as two timelines:

1. **A timeline for regular development activities**. We aim to operate the ENVRI knowledge base in a continuous integration and deployment manner. There will be a major release every three months in the operation version.
2. **A timeline for agile tasks** driven by dynamic needs and priorities from the community. The typical duration of an agile task will last 3-4 months.

5 References

[1] CERIF: https://www.eurocris.org/services/main-features-cerif

[2] SEMAF https://zenodo.org/record/4651421#.YGSd2y0RonU

[3] TF1 White Paper: https://docs.google.com/document/d/1RiCTzqqmOGvotJHzKhfEqyFAxjSHuv3110Ow9BDov-Q/edit

[4] TF2 White Paper https://docs.google.com/document/d/1cXfEzspbejGJPz_fCY-XPO0OjYF_as0S/edit?rtpof=true

[5] Schultes E., Magagna B., Hettne K.M., Pergl R., Suchánek M., Kuhn T. (2020) Reusable FAIR Implementation Profiles as Accelerators of FAIR Convergence. In: Grossmann G., Ram S. (eds) Advances in Conceptual Modeling. ER 2020. Lecture Notes in Computer Science, vol 12584. Springer, Cham. https://doi.org/10.1007/978-3-030-65847-2_13

[6] S Koulouzis, Y Shi, Y Wan, R Bianchi, D Kissling, Z Zhao, Enabling "LiDAR data processing" as a service in a Jupyter environmen, EGU21 Poster

[7] Zhao, Z., Liao, X., Martin, P., Maduro, J., Thijsse, P., Schaap, D., Stocker, M., Goldfarb, D., Magagna, B.: Knowledge-as-a-Service: A Community Knowledge Base for Research Infrastructures in Environmental and Earth Sciences. In: 2019 IEEE World Congress on Services (SERVICES). pp. 127–132. IEEE, Milan, Italy (2019). https://doi.org/10.1109/SERVICES.2019.00041

[8] Matthew B. Jones, Margaret O'Brien, Bryce Mecum, Carl Boettiger, Mark Schildhauer, Mitchell Maier, Timothy Whiteaker, Stevan Earl, Steven Chong. 2019. Ecological Metadata Language version 2.2.0. KNB Data Repository. doi:10.5063/F11834T2

6 Appendix 1: Glossary

ACTRIS	Aerosols, Clouds, and Trace gases Research InfraStructure network
Catalogue (Metadata)	A collection of metadata, usually established to make the metadata available to a community. A metadata catalogue has an access service.
CERIF	Common European Research Information Format
CODATA	Committee on data for Science and Technology
DOM	Document Object Model (DOM) is the data representation of the objects that comprise the structure and content of a document on the web.
ENVRI	(1) The ENVRI Community of Environmental Research Infrastructures. (2) FP7 project on Implementation of common solutions for a cluster of ESFRI infrastructures in the field of Environmental Sciences.
ENVRIplus	ENVRIplus is a Horizon 2020 project bringing together Environmental and Earth System Research Infrastructures, projects, and networks together with technical specialist partners to create a more coherent, interdisciplinary, and interoperable cluster of Environmental Research Infrastructures across Europe.
ENVRI-FAIR	An EU-funded project which stands for ENVironmental Research Infrastructures building Fair services Accessible for society, Innovation, and Research.
FAIR	Findability, Accessibility, Interoperability, and Reusability of digital assets
Elastic Search	Elasticsearch is a search engine based on the Lucene library.
EOSC	European Open Science Cloud
FITSM	The name for a family of standards for lightweight IT service management (ITSM).
GDPR	General Data Protection Regulation, a regulation in EU law on data protection and privacy in the European Union and the European Economic Area.
GO FAIR	A bottom-up international approach for the practical implementation of the European Open Science Cloud (EOSC).
GUI	A GUI (graphical user interface) is a system of interactive visual components for computer software.
H2020	Horizon 2020, European level research funding scheme
Knowledge Base (KB)	(1) A store of information or data that is available to draw on. (2) The underlying set of facts, assumptions, and rules which a computer system has available to solve a problem.
LifeWatch	European e-Science infrastructure for biodiversity and ecosystem research
Metadata	Data that describes other data. Metadata summarizes basic information about data, which can make finding and working with particular instances of data easier.
NetCDF	A file format
RM-ODP	Reference Model of Open Distributed Processing (RM-ODP) is a reference model in computer science, which provides a coordinating framework for the standardisation of open distributed processing (ODP)
OIL-e	Ontology of the ENVRI Reference Model

Ontology	(In computer science and information science) an ontology is a formal naming and definition of the types, properties, and interrelationships of the entities that really or fundamentally exist for a particular domain of discourse.
Ontowiki	A free and open-source semantic wiki application, meant to serve as an ontology editor and a knowledge acquisition system.
Open Semantic Search	A free software for building own Search Engine, an explorer for discovery of large document collections, media monitoring, text analytics, document analysis & text mining platform based on Apache Solr or Elasticsearch.
OWL	Web Ontology language
Provenance	The pathway of data generation from raw data to the actual state of data
RDA	Research Data Alliance
RDBMS	A software system used to maintain relational databases
RDF	Resource Description Framework
RI	Research Infrastructure
SPARQL	SPARQL is an RDF query language—that is, a semantic query language for databases—able to retrieve and manipulate data stored in Resource Description Framework format.
Semantics	The encoding of meaning using a formal language.
Semantic Mediawiki	Semantic MediaWiki is an extension to MediaWiki that allows for annotating semantic data within wiki pages, thus turning a wiki that incorporates the extension into a semantic wiki.
Triple	A triple is a data entity composed of subject-predicate-object
Triplestores	A triplestore or RDF store is a purpose-built database for the storage and retrieval of triples through semantic queries.
VRE	virtual research environment
Wikidata	A collaboratively edited multilingual knowledge graph hosted by the Wikimedia Foundation.
W3C	World Wide Web Consortium
WP	Work Package
YAML	A human-readable data-serialization language.