D10.2
Roadmap for implementation of FAIR concepts

Work Package	WP10
Lead partner	INGV
Status	Final
Deliverable type	Report
Dissemination level	Public
Due date	30 June 2020
Submission date	29 June 2020

Deliverable abstract

D10.2 “Roadmap for implementation of FAIR concepts” reports about the agreed strategies and roadmap for implementing FAIR principles in the Solid Earth domain community. FAIR principles have the merit of creating a common background of knowledge to engage communities in providing data in a standard way thus easing interoperability and data sharing. However, they do not explicitly refer neither to methodologies nor actual technical activities that RIs might adopt in order to implement software and technologies actually delivering FAIR data.

In this deliverable we therefore describe the methodology we used to move from principles to technical activities (chapter 2), then we describe the common work done in the whole WP10 related to the FAIR assessment and planning process (chapter 3); finally, for each task, we describe the roadmap in terms of gap analysis and emerging technical activities to be undertaken to fill the gaps (chapter 4).

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	Daniele Bailo	INGV	01.06.2020
Contributing Authors	Keith Jeffery Luca Trani Jean-Baptiste Roquencourt Tor Langeland Ivan Rodero Michele Manunta	UKRI (BGS) KNMI BRGM NORCE (UiB) EMSO CNR (IREA)	
Reviewer(s)	As above		
Approver	Andreas Petzold	FZJ	29.06.2020

DELIVERY LOG

Issue	Date	Comment	Author
V 0.1	22.06.2020	Final version submitted to project management	Daniele Bailo

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 main goal of this deliverable is to report about the agreed strategies and roadmap for implementing FAIR principles in the Solid Earth domain community. FAIR principles have the merit of creating a common background of knowledge to engage communities in providing data in a standard way thus easing interoperability and data sharing. However, they do not explicitly refer to methodologies nor actual technical activities that RIs might adopt in order to implement software and technologies actually delivering FAIR data.
In this deliverable we therefore describe the methodology we used to move from principles to technical activities (chapter 2), then we describe the common work done in the whole WP10 related to the FAIR assessment and planning process (chapter 3); finally, for each task, we describe the roadmap in terms of gap analysis and emerging technical activities to be undertaken to fill the gaps (chapter 4).

2 Methodology

Research Infrastructure (RIs) implementers setting up or upgrading an existing RI usually follow a system development life-cycle (SDLC) process (Blanchard, 2004). In the current deliverable, an SDLC inspired by the waterfall development model and by a previous work (Bailo, 2020) was proposed, which encompasses the following steps: a) analysis, including use cases and requirements collection; b) design, including architecture design and identification of architectural components matching requirements, c) implementation, through software developments and adoption of suitable technologies, d) test, e) operation and maintenance. Only steps from a) to d) are considered in the current deliverable, as not all the services are committed to be operational.
Although FAIR principles claim to be technically non-prescriptive (Mons, B., 2017), they need to intersect and be part of SDLC since its early phases, leading to three main questions:

I. where FAIR principles have to be considered in the SDLC
II. what FAIR detailed principles should be implemented first, and what would be a correct sequence and time-line
III. how FAIR principles should be technically addressed

FAIR principles define what the system should provide and how it should be provided; they can therefore be considered as requirements in the SDLC and should be taken into account during the analysis phase.
However, how to manage contexts where a Research Infrastructure already exists and needs to be upgraded to be compliant to FAIR principles is still an open question. This is indeed a common status for many RIs, and most importantly for those represented by tasks 10.3 to 10.6 in Work Package 10.
The reorganization of the FAIR detailed principles into a four-stages roadmap together with potential technical activities to implement them at each of the stages of the roadmap described in (Bailo, 2020) may provide a perspective to answer to this latter question, but wasn't used in the current methodology, that was more inclined to adopt a lean a easy to apply process.
Therefore, the methodology used to enhance RIs services represented by tasks 10.3 to 10.7 (Fig 1) was simplified as follows:

1. FAIRness assessment & Gap Analysis. In order to enhance RIs services, its compliance to the detailed FAIR principles needs to be evaluated and a gap-analysis performed to define technical activities addressing criticalities emerged during the evaluation. Currently, several initiatives are devoted to evaluating compliance to FAIR principles[1] (Wilkinson, 2019); as consensus about one or more evaluation methods is reached within the scientific community, they might be adopted as canonical evaluations. As for the current WP10 work, a simple approach based on discussion based evaluation for each of the FAIR principles applied to a specific RI service, was applied, as shown in table 1A and Table 1B.
2. Technical activities definition. On the basis of the FAIRness evaluation results and the emerged gaps, a first draft of the actual technical activities needed to fulfil a specific FAIR principle were defined in the domain specific context by following steps from design (b) to operation (e) of the SLDC.

[1] http://blog.ukdataservice.ac.uk/fair-data-assessment-tool/
https://www.biorxiv.org/content/biorxiv/early/2018/09/25/418376.full.pdf
https://www.go-fair.org/2017/12/11/metrics-evaluation-fairness/ and http://aims.fao.org/activity/blog/put-fair-principles-practice-and-enjoy-your-data (all accessed 05.06.2020)

In this two-steps FAIR adoption process the SDLC is not confined to the technical implementation step (3), as its analysis phase encompasses also the Definition of the stages to implement (1) and FAIRness evaluation (2) steps, which is where FAIR detailed principles are defined and elicited as requirements (Fig. 1).

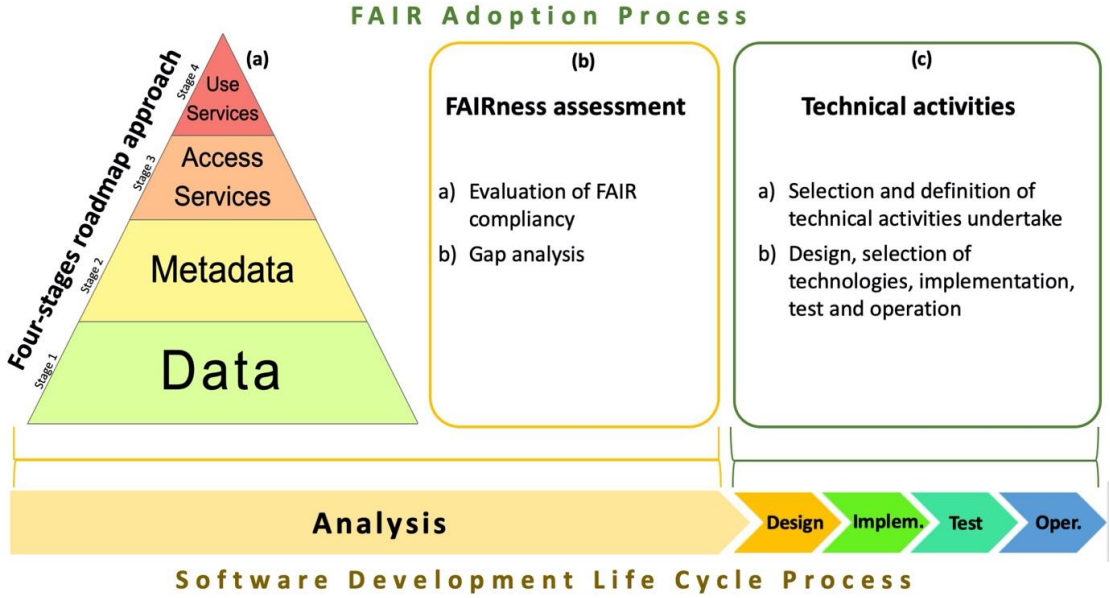

Fig 1 - FAIR adoption and Software Development Life Cycle processes are shown in parallel and correspondences between the two are highlighted, taken from (Bailo, 2020). Only steps (b) and (c) were applied in the current work in WP10: the second step (b) was devoted to the evaluation FAIR principles against services (WP tasks) and to a gap analysis to make emerge the priority by which FAIR principles need to be addressed; the third step (c) is devoted, on the basis of the gap analysis, to the selection of technical activities to undertake, and then to actual implementation.

Software Development Life Cycle is here compared to the FAIR adoption process. What emerges is that step (b) is carried out in the Analysis Phase in the SDLC; indeed, FAIR principles are elicited as requirement and gap analysis in order to understand which ones will be implemented at each "round" of implementations performed. Phases from design to operation are executed in the "technical activities" step of the FAIR adoption process.
The process described by the above methodology was agreed by all partners in Work Package 10, with an important milestone at the ENVRI week in Dresden, Germany, 3-7 February 2020, where the process was discussed, updated, agreed and kicked-off as described in the next section.

3 FAIR methodology implementation

The implementation of the methodology described above therefore relies on two main activities:
1. FAIRness assessment & Gap Analysis.
2. Technical activities definition

Such activities were carried out:
a) in a horizontal cooperative way, by intra-Tasks face to face discussion at the various ENVRI-FAIR meetings (e.g. ENVRI week in Dresden, Germany, 3-7 February 2020) and at the periodic teleconferences call
b) in a vertical way, within tasks, where the work was organized internally to undertake assessment, gap analysis and activities definition
All tasks undertook this work, except T10.7 where the services belonging to the EMSO distributed RIs had a different approach, as described in section "Task 10.7 - Marine (EMSO" of the current deliverable.
The work carried out is summarized in FAIRness exercise (Table 1A and 1B) below.

Table 1A: excerpt of the "FAIRness exercise" worksheet, "FAIR assessment" spreadsheet. Full worksheet available as shared document[2]

	F1	F2	F3
SEISMOLOGY			
Primary Seismic waveform data	**data:** only some datacenters, assign EPIC PIDs to daily waveforms (GFZ, KNMI, INGV, NOA), **metadata:** when they are available they are global, **ICS-C:** capabilty of CERIF supporting it, but not implemented	yes	Yes, when EPIC PID is available
SATELLITE			
Line of sight displacement time series	**data:** NO, **metadata:** YES but not global	**data:** ISO19115 (YES)	**data**: referenced by URL (not a PID)
ICS-D			
SWIRRL Juputer notebook as an instance (Metadata Handled by SWIRRL)	Entities with metadata describing a Jupyter Instances are uniquely identified and persisted	Ingested data into Jupyter through SWIRRL gets a unique identifier, which is generated by SWIRRL The var:Jupyter resource of the following template shows the current set of metadata https://openprovenance.org/store/documents/2297	metadata include the identifier of the Jupyter Resource.

Table 1B: excerpt of the "FAIRness exercise" worksheet, "Emerging Activities" spreadsheet. Full worksheet available as shared document[3]

	F1	F2	F3
SEISMOLOGY			
Primary Seismic waveform data	Promote the use of PIDs in EIDA data centres		
SATELLITE			

[2] Full table : https://docs.google.com/spreadsheets/d/1A1qDdznwDqdcmGi5vhgjRJzpYqk1aL_MXDZywRM72MA/edit#gid=1979957505 (accessed 05.06.2020)

[3] Full table : https://docs.google.com/spreadsheets/d/1A1qDdznwDqdcmGi5vhgjRJzpYqk1aL_MXDZywRM72MA/edit#gid=1979957505 (accessed 05.06.2020)

Line of sight displacement time series	Activities to use a PID system have to be put in place. The community has never adopted any standard. Specific experiments have been carried out with Zenodo. TCS is evaluating which system is more suitable. The lack of PID impacts also on other FAIR principles		
ICS-D			
SWIRRL API - webservice - API Gateway	BRGM is working on providing an API Gateway to access services		

On each row of the above two tables, services from the RIs represented by tasks in WP10 are discussed against the FAIR principles marked in each columns.

Table 1A reports the FAIR assessment: the evaluation is carried out in a descriptive way. Other assessment methodologies may be considered in the future, as the scientific community seems to be more mature with respect to the convergence of a common FAIR metrics or approach to evaluation, as exemplified by the work carried out, for instance, in the context of FAIR Data Maturity Model WG[4].

Table 1B reports the Emerging activities, that is to say activities that are needed to fill in the gap evidenced in Table 1A.

4 Gap analysis and Activities roadmap for RI services

In the current section, we report the application of the methodology described above, together with all other relevant elements and discussions carried out in the Tasks representing RI services, i.e. Tasks from 10.3 to 10.7. Gaps, potential improvements and activities are discussed in a detailed way.

4.1 Task 10.3 - ICS-C

EPOS has been driven with FAIRness in mind from the beginning and, in this sense, is already FAIR to a certain degree.
However as FAIRness is a journey, this task aims to improve the FAIRness of EPOS central component known as the ICS-C.
Following the SDLC, the tasks start with sharing a common approach of the FAIR principles between the ICS-C stakeholders.
From then on, a gap analysis will follow. As EPOS ICS-C has already boarded the FAIR journey, the gap analysis will help us on refining the FAIR activities and their prioritization.
These activities can be split between in layers:

- From data providers' perspective, aka TCS: increase their data (DDSS) visibility and usability, providing PIDs.
- From the ICS-C itself : improve the process of FAIRness, enhanced AAAI integration
- From the consumers perspective : richer metadata catalogue
- From the reusability perspective : link to EOSC and the ICS-D

Hence the FAIR roadmap of the task 10.3 involved actions which require engaging multiple stakeholders with the relevant skills assets.

[4] FAIR Data Maturity Model WG web page: https://www.rd-alliance.org/groups/fair-data-maturity-model-wg (accessed 05.06.2020)

4.1.1 FAIR assessment and gap analysis

As FAIR is still improving, so does the FAIR assessment. The FAIR assessment method has to be defined by the assessors. Therefore a clear comprehension of the FAIR principles is mandatory.
The FAIR principles sometimes provide examples to guide the assessors. However it is not always the case, because giving examples can narrow down the scope of the FAIR principle assessed. Therefore a common approach and understanding of the FAIR principle between the team is mandatory.
The common understanding of the principles was reached by holding multiple face to face meetings and web conferences.
They involved different stakeholders: FAIR experts, technical experts from task 10.3 to 10.7.
During these meetings we worked on practical examples, as this has been the method in the EPOS community for a long time. By working on the practical example, we narrow the scope at first to facilitate the building of a common approach. Once down for every task we can broaden the scope.
This approach, based on the Agile Principles, has proven to be successful in EPOS. It allows us to have an early Minimum Viable Product and strengthens the community bonds and culture.

4.1.1.1 FAIR assessment

In task 10.3 we have to consider the full EPOS landscape (Fig.2), and then focus on the parts that are only related to the ICS-C.

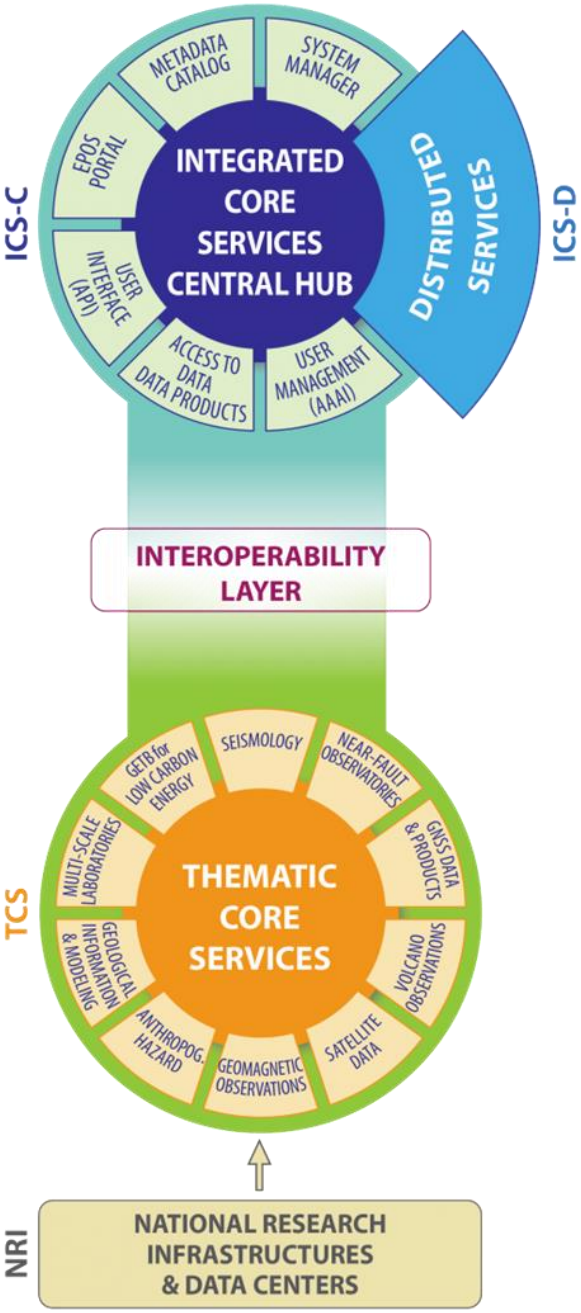

Fig. 2: The EPOS architecture has been designed to organize and manage the interactions among different EPOS actors and assets. Three complementary elements form the ICT architecture: The Integrated Core Services (ICS), The Thematic Core Services (TCS), The National Research Infrastructures (NRIs), technical interface between TCS and ICS is the interoperability layer, which guarantees communication and interoperability.

First we need to assess what we call metadata and data. It depends on the point of view, if it is input or output.
Therefore we need to review the metadata/data workflow:

1. TCS are publishing DDSS.
2. TCS are providing metadata in the EPOS-DCAT-AP model for their DDSS (Web API, publications, raw data, etc.)

3. metadata in the EPOS-DCAT-AP model are ingested in CERIF model
4. CERIF model is consumed by user or machines via an API
5. The API can be consumed via EPOS ICS-C GUI.

Hence we could come with the following distinction between metadata and data in the ICS-C context.
1. The TCS are providing DDSS, they are data.
2. The ICS-C catalogue contains metadata of the TCS (aka data providers). So from the TCS perspective, the metadata catalogue contains metadata.
3. From the API perspective, TCS metadata are still metadata, but how about the ICS-C itself

In the FAIR assessment of the ICS-C, we will then focus on the ICS-C metadata only and on step 2 to 4 on the data/metadata workflow.
Enlighten with that vision, we can narrow it down to the following technical component of the ICS-C:

- What we can call the visibility layer, how EPOS ICS-C is visible for external user :
  - the ICS-C API
  - the ISC-C GUI
- What we can call the model layer: the EPOS data models
  - EPOS-DCAT-AP
  - CERIF

4.1.2 Gap Analysis

Findable
The EPOS ICS-C is the infrastructure delivering Earth Science dataset, data product, software and service (DDSS). In order to FIND the DDSS, they first have to be registered in the Metadata Catalogue, and made FINDable within the Metadata Catalogue entry point, which is the ICS-C API.
The ICS-C API provides an endpoint for FINDing data. This endpoint is named "discovery" and provides 2 methods. The one that allows a client (User or Machine) to search a resource, aka DDSS is "/resources/search", and return DDSS with their Identifiers.

Thus EPOS already provides a way to FIND resources registered in the ICS-C Catalog based on a search over attributes and returning metadata records with IDs.

We have seen earlier that EPOS already provides a consistent way of finding DDSS. However we identified some improvement.
The Metadata Catalogue is based on the CERIF model. The CERIF model recommends using UUID. In the actual implementation of the CERIF model within the ICS-C, IDs describing the DDSS already exist. Those IDs are Universal and Unique in EPOS. If we want to have to ID Unique and Universal on the web, we need to provide a UUID in the form of URI.
The other point of improvement is the format of the answer from the ICS-C API. In the case of an existing external registry and search engine, EPOS ICS-C catalogue would be better indexed if the answers were compatible with JSON-LD.

Accessible
The metadata access is guaranteed through a standard protocol (HTTP) based on their references. An access control is possible under this protocol. However, no authorisation control is provided yet by the catalog. This is an area of current work. The ICS-C implementation guarantees the sustainability of the metadata even if the data is inaccessible or missing at some point. The sustainability is achieved a)technically with a CERIF model database and its API accessible through 2 hosting organisations facilities (BGS and BRGM) and b) legally, by the EPOS ERIC.

Interoperable
The CERIF model provides the interoperable capabilities. The semantic layer which stores vocabularies imported from FAIR compliant vocabularies. The database implementation ensures the qualified reference between metadata elements. The ICS-C has the basement for interoperability on the storage layer.
The dissemination layer, the ICS-C API, is a homemade open REST API, because it is both used for research, displaying and workspace management. We will focus on the research endpoint "/resources": its function is to answer user queries. Therefore the query endpoint and the answer needed to be assessed. The ICS-C API, as written earlier, provides the discovery endpoint. Although it is open it

would be worth investigating Opensearch URL for reaching a higher interoperability level on the query flow. It would allow users to add directly in their browser the ICS-C catalog. The answer content is a geo/json format which is a well-known format and as such presents an interesting level in terms of interoperability. Here it would be interesting to improve it with a content negotiation pattern which would return either: CERIF/XML, RDF/XML or EPOS-DCAT-AP/JSON.

Reusable
The CERIF model provides a basis for rich metadata representation. Indeed it includes a considerable amount of attributes allowing the data to be well described, and thus easily understandable. This allows for users (human or machine) to decide whether to use the data or not and eventually combine it with other data. Nonetheless, some aspects that are easily integrable to the current CERIF implementation and could enhance the reusability of data are not taken into account yet. For instance, CERIF and EPOS-DCAT-AP provide mechanisms for referring to licenses but not populated fully. The same goes for explicit provenance information that are missing though CERIF natively provides provenance information by the temporal attributes on the links between base objects. In addition, a work is currently held to convert from/to PROV information to/from CERIF. Although, CERIF is a EU recommended model, the output provided by the Web API is not formalized following a semantic/ technical standards such as : EPOS-DCAT-AP, CERIF, JSON-LD.

4.1.3 Technical Activities and Implementation

As the EPOS community had in mind from the very start the FAIR principles, and have organized its collaboration on continuous improvement of the FAIRness of the ICS-C, the ICS-C is deemed to be FAIR enough. Therefore the FAIR assessment has helped in paving the next stage of FAIRness. It has also provided good intel on how we will prioritize the activities regarding EPOS Strategy Plan.
The roadmap is then the following:

- Defining the URI pattern for reaching specific DDSS metadata elements, this would provide PIDs.
- Implementing the URI pattern in the EPOS ICS-C API.
- Implementing content negotiation for the EPOS ICS-C API answer format.
- Research the potential of search APIs.
- In link with the other communities, and in order to facilitate cross-domain search, providing a consistent way of managing vocabularies that would prevent communities to redo the work done by others. This boils down to choosing the right tools and defining the policies of managing vocabularies: namespace, workflow and responsibilities on vocabularies. Being pragmatic we will reuse the one used in the TCS geology community (http://data.geoscience.earth/ncl/)
- Extending EPOS-DCAT-AP and CERIF model implementation to support the entities Facility and Equipment. We intend to demonstrate this extension with EPOS communities involved in ENVR-IFAIR.

4.2 Task 10.4 - ICS-D

Task 10.4 Implementation of FAIR roadmap in ICS-D services demonstrates a virtual research environment (VRE) concept as an ICS-D implementation under the lead of UiB/NORCE in the context of EPOS. The example cases that will be used will focus on a visualization and analysis web platform which can be launched from a data-discovery session, with access to the workspace generated in this session. This will thus contribute to EOSC of data and services where software applications can be routed to relevant data and vice versa.

Integration and interoperability of the visualization and analysis web platform with the ICS-C have been conducted with special emphasis on the FAIR principles. A demonstrator of the VRE concept will be developed focusing on visualization and analysis platforms integrated with ICS-C. The goals of this task are thus to implement technical elements for ICS-D to achieve FAIRness maturity and EOSC compatibility. This will include implementation of appropriate AAAI, entries in the EPOS rich metadata service catalogue compatible with (or convertible to) EOSC service catalogue. We also validate and test a set of use cases for EOSC interoperability through ICS-C.

The ICS-D implementation in this project requires the following components, treated as “services” in table 1A and 1B, to be in place:

- SWIRRL API – webservice. This is a framework used for executing workflows for staging required data and for starting the specific ICS-D, in this project specifically Enlighten-web and Jupyter notebook .
- Enlighten-web – This is a web application for interactive visual analysis. In this project we will implement a visualization workflow that prepares data from the ICS-C workspace and that uses Enlighten-web for visualizing the data.
- Jupyter notebook – This is a web programming environment for analysis. It will be made available in the same manner as described for Enlighten-web.
- Hosting facility - Deployment of the ICS-D services at BGRM.

The work so far has focused on FAIR assessments of the ICS-D components and technical work for getting the components up and running.

4.2.1 FAIR assessment and gap analysis

We have applied a bottom-up approach starting with analysing the FAIRness of the components of the ICS-D. The FAIRness of the ICS-D as a whole will not depend only on the FAIRness of the components, but also on the interface to and interaction with the ICS-C, i.e. how the user interacts with ICS-C to access data and initiates ICS-D services using the data in the ICS-C workspace.

4.2.1.1 SWIRRL API, Jupyter notebook

SWIRRL offers a service which allows the production of new analysis, methods and data. As such its FAIRness is two folds:

1 – As a SWIRRL API service. It should be included into the ICS-C catalogue and described by a collection of metadata characterising its offering. These should include UUID, documentation (Git Repository), description, type of services offered (Visualisation, Development), target communities, keywords, all together addressing the F of the FAIR principles. Moreover, it should specify the access endpoints via a machine-understandable descriptor (Which addresses the A and I of the FAIR). More instances of SWIRRL might exist but we foresee a single instance serving EPOS. SWIRRL is not meant to be used by users directly, but it serves Portals, such as Science Gateways. Being SWIRRL a service, the R part should be addressed by specifying the right version of the software currently deployed and used by EPOS.
2 – As a SWIRRL tool instance (Jupyter or Enlighten Instance, Data Staging workflow). SWIRRL is used to spawn instances of Jupyter Notebooks, Enlighten-web or workflow jobs. The former is described and catalogued automatically by SWIRRL with a set of metadata (F) which are included in a larger data model adopting provenance standards (R). Activities such as instance creation, update, and execution of workflow are recorded in terms of a collection of metadata and placed within a provenance model. For instance, Fig. 3 shows the template for the creation of a new Notebook instance, while Fig. 4 shows a template which represents the provenance of performing an update of libraries used within a notebook environment. Workflow executions and their results are also recorded in a provenance model. See template in Fig. 5. The newly produced data files are thereby described and linked with the generating process and environment. The information is placed within a graph database and exposed via a Websevice API (AI).

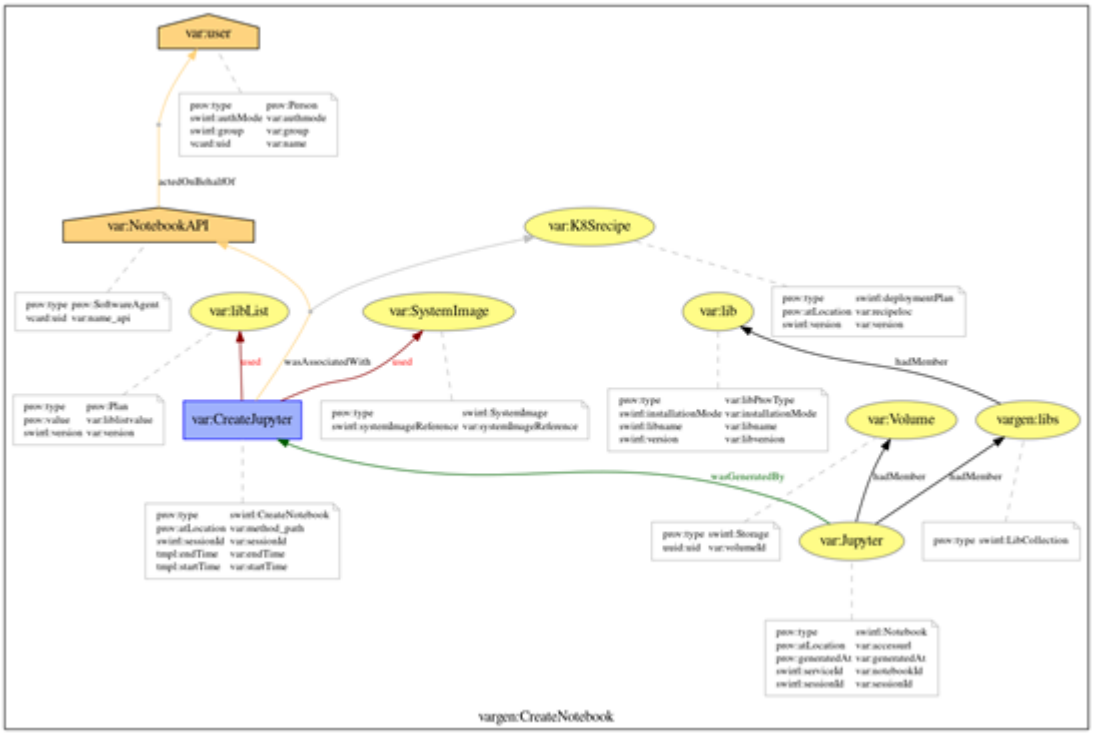

Fig. 3. Creating a Notebook Instance. The provenance graph includes the metadata and provenance relationships associated with the creation of a new Notebook Instance.

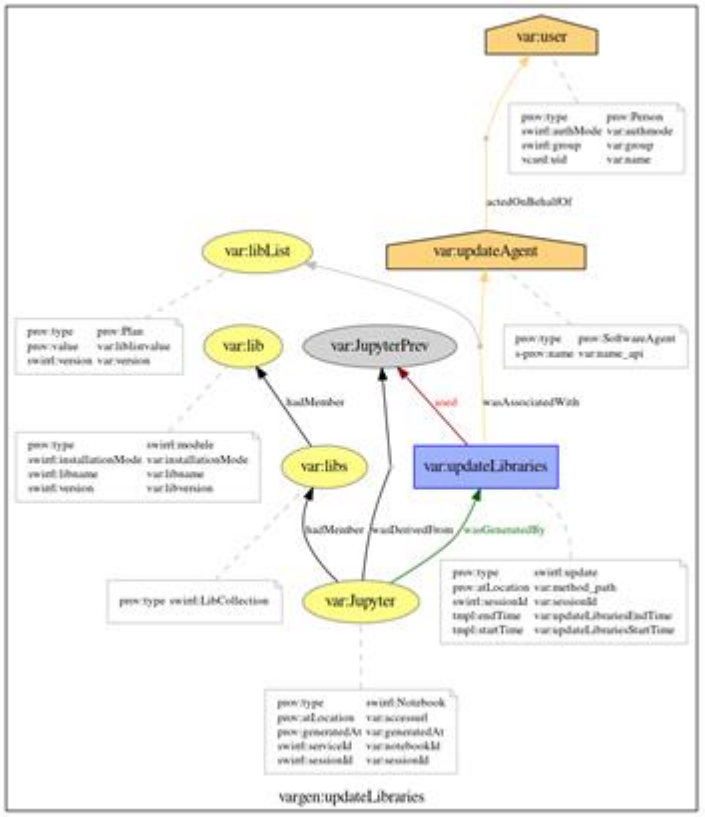

Fig. 4. Notebook Update Template.

Future work would address the generation and instantiation of templates for the activities related to the Enlighten-web service. We foresee that many of the current templates could be re-used or adapted. Moreover, a new template can be designed aiming at achieving specific FAIR objectives for this tool.

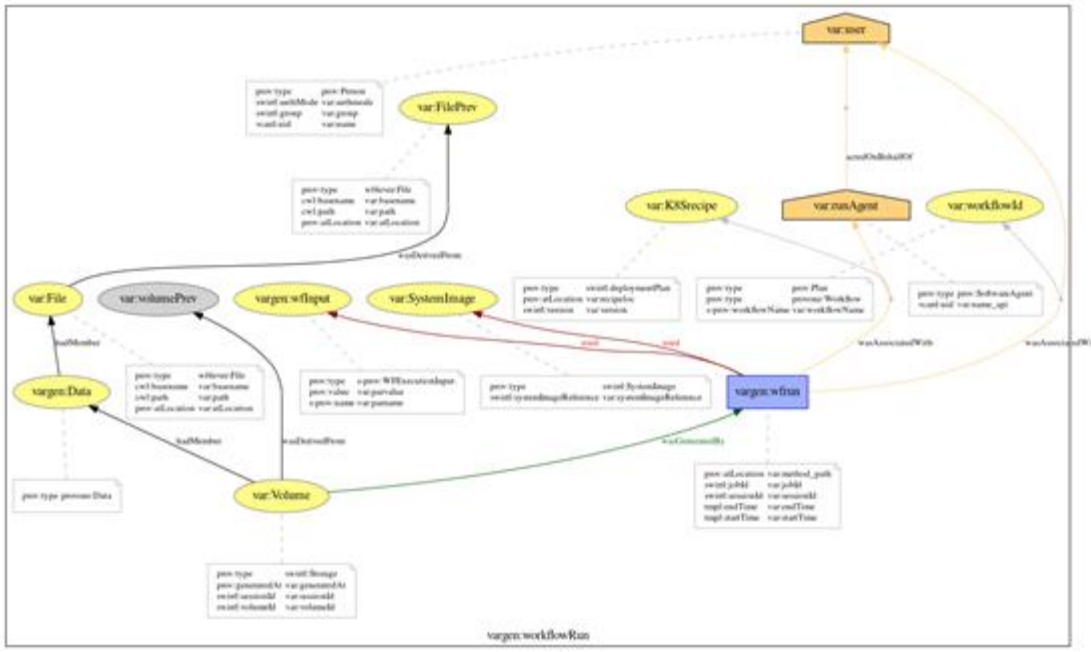

Fig. 5. Provenance Template Describing the execution of a Computational Workflow in SWIRRL.

4.2.1.2 Enlighten-web

As for the SWIRRL API, the FAIRness of Enlighten-web in this context is two-fold:
1 – Enlighten-web as a Service. This refers to an entry in the ICS-C Catalogue with metadata describing the Enlighten-web visualization tool. Here we have identified the major gap to be that currently there is no metadata defined for Enlighten-web. We need to decide on vocabulary to use and define an EPOS-DCAT-AP description that can be ingested into the ICS-C catalogue. Other gaps with regards to how these metadata can be retrieved and accessed will depend on how the metadata are registered and managed in the ICS-C catalogue.

2 – Enlighten-web visualization as an instance. Here, gaps with regards to required identifiers and findability can be filled by applying solutions along the same lines as has been done for Jupyter Notebook instances. There is also a need for defining metadata for the actual visualizations. This can be achieved by extracting data from Enlighten-web native plot specifications in json format and translate it to relevant vocabularies in EPOS-DCAT-AP.

4.2.2 Technical Activities and Implementation

We have performed a preliminary integration of metadata management for a particular ICS-D. The work has included real adoption of provenance technologies to handle description and evolution of services, as well as the execution of data-management (e.g. staging and pre-processing) workflows. We have realized workflows involving Jupyter Notebooks and performed tests that indicate that Enlighten-web can be integrated in a similar manner.
Based on the FAIR assessment and gap analysis (Tables 1A and 1B) described in section 2 and 3, we have identified the following main topics for remaining technical activities and implementation:

- Metadata must be defined for the Jupyter Notebook as a service and Enlighten-web visualization service. The metadata must be described in EPOS-DCAT-AP ttl files. We must compare our requirements with existing EPOS-DCAT-AP and identify gaps. When EPOS-

DCAT-AP has been updated to fill identified gaps, the defined metadata can be ingested into CERIF.

- For SWIRRL Jupyter notebook as an instance and Enlighten-web as an instance, the protocol does not yet allow for an authentication and authorisation procedure, but OAUTH is envisaged also with delegation.
- Define metadata for Enlighten-web instances. These metadata describe the visualizations created using Enlighten-web. The metadata will be created when the user stores visualization. For the generation and instantiation of templates for the activities related to the Enlighten-web service, we foresee that many of the templates used for Jupyter Notebook instances can be re-used or adapted. Moreover, new templates can be designed aiming at achieving specific FAIR objectives for Enlighten-web instances.
- We must implement microservices for converting data from ICS-C workspace to Enlighten-web input data. We foresee that ICS-C workspace contents will be in the form of URLs defining searches for data. We do not want to perform these searches every time the user accesses a SWIRRL Enlighten-web instance.
- Licensing is also a remaining issue for most of the ICS-D components.
- We will pursue integration of the ICS-D within the EPOS Portal (ICS-C). For the FAIR principles F4, A1, A1.1 and A1.2, compliance will depend on the way ICS-D services are made available from ICS-C.
- The activities concerning the FAIRness of service instances are fundamental to collect an initial baseline of metadata that can be used by SWIRRL to implement new API methods offering restoring actions. These aim to reset the Notebook or an Enlighten instance to a previous state, and to reproduce the environment and the data generated with these tools by peers. This would demonstrate and actual adoption of the FAIR capabilities of the tool in real use-cases.

Filling all these gaps requires a substantial amount of work that exceeds the available budget. We will therefore carefully prioritize the tasks to secure as optimal usage of the available budget as possible.

4.3 Task 10.5 - Seismology

This task focuses on the application of the FAIR principles to seismological data and products. It will develop awareness and share knowledge about FAIR in the seismological domain and research effective approaches to help establish the FAIR principles into existing practices and methods. The task will build on the results of previous and ongoing activities carried out in the context of projects such as EUDAT2020, EOSC-Pilot, EOSC-hub, VERCE, EPOS-IP and DARE.
The aim is to produce a framework to address the FAIR aspects in the different phases of the data lifecycle covering acquisition, curation, computation, dissemination and publication. The goals of this task are:

- Integrate of data management policies, data descriptions (metadata and quality control), and product descriptions (metadata, provenance, methods, policies).
- Harmonisation of results of activities in EOSC-hub (data lifecycle, curation and products generation/containerised), VERCE (Science Gateway and cross-platform provenance-aware seismological workflows), DARE (Data-intensive research on the Cloud and large-scale lineage services), where virtualisation and containerisation will constitute the new underlying computational ecosystem.
- Assessment in cooperation with T10.3 (ICS-C) and task 10.4 (ICS-D) of the metadata portfolios and software components that will deliver new FAIR research products through virtualised appliances, with special concerns on their reproducibility. The solutions will be harmonised across the subdomain's VREs.
- Policies for the FAIR Dissemination of experimental outputs (including controlled authorisation mechanisms).

4.3.1 FAIR assessment and gap analysis

In this first part of the project we performed a FAIR assessment targeting a specific seismological product: seismic waveform data. We started with an analysis of the current status of seismic waveform data, metadata, data services and tools. By adopting the methodology described in Section 2, we addressed systematically aspects of the FAIR principles and for each one we: a) set feasible goals for

the targeted level of FAIRness, b) identified the related gaps and c) defined activities to bridge those gaps.
In the following sections we summarise and discuss the results of our analysis by addressing the different dimensions of FAIR.

Findable
Our initial focus was on the find-ability of seismic waveform data. In particular, we considered the current adoption of Persistent Identifiers (PIDs) associated with seismic waveforms in the context of the ORFEUS-EIDA community. We set as a goal to reach the broadest possible adoption of PIDs within ORFEUS-EIDA data centres. Ideally, we would aim at an harmonised and consistent way to associate and manage PIDs across the whole ORFEUS-EIDA. The current landscape includes data centres that have been pioneering PID solutions in projects such as EUDAT and EOSC-hub. Those delivered in some cases operational products and services. In particular, the combination of EPIC Handle and B2SAFE were successfully adopted for minting PIDs and achieving long-term preservation of large seismic waveform archives. However, the impact of such an integrated solution on the operational infrastructure of some data centres is considered too heavy in terms of resources and capacity required. For this reason, a more flexible and lightweight solution is desirable. For instance, in some cases data centres expressed the need to define and manage their own identifiers. A discussion has been initiated and will be continued in order to achieve the highest level of harmonisation possible but at the same time taking into account the specific needs. A metadata catalogue has been identified as a key component to achieve such a goal, namely WFCatalog. WFCatalog has become an ORFEUS-EIDA standard and it is operated in all major data centres in Europe.
WFCatalog contains detailed and rich descriptions that enable discovery of and support access to seismic waveform data. The WFCatalog data model includes PIDs as metadata features; however, as PIDs are not adopted and implemented by all the ORFEUS-EIDA data centres they are often not populated in the catalogue. WFCatalog offers functionalities to cover most of the findability aspects, activities are required to implement those functionalities in an harmonised way across ORFEUS-EIDA.

Accessible
At present seismic waveforms are made available and delivered to users via community data services that adhere to international standards (e.g. FDSN). ORFEUS-EIDA offers tools, portals and catalogues to enable interactive and/or automated discovery and access. However, some adjustments and extensions are required in the current access mechanisms in order to increase the FAIRness level e.g. by supporting PID-based queries and by defining and establishing policies for long-term metadata management.

Interoperable
The interoperability aspect requires work on the definition and establishment of a common seismological vocabulary. This activity has been initiated at a broader level in the EPOS-IP project by engaging representatives of the community and by providing them with a framework for collaboration and knowledge sharing in order to discuss and reach agreements on definitions of concepts. Also, a mechanism to represent such agreed definitions has been provided by adopting SKOS and Linked Data integrated in EPOS-DCAT-AP. However, the establishment of a common vocabulary is a long-term, ongoing activity that requires clear processes and governance to oversee and manage authoritative definitions. In this context our goal is to continue the work initiated and to achieve a first set of agreed definitions that could be made available in the current seismic metadata catalogue and services. Also in this case by extending WFCatalog and its API we plan to reach an increased level of FAIRness.

Reusable
With respect to reusability our goal is the integration of computation and data supported by rich provenance that would eventually enable reproducibility. When considering the current status of the community this is a quite ambitious but feasible goal. It builds on activities initiated in projects such as EOSC-hub, EPOS-IP, DARE [refs] and continued in task 10.4. In particular we are considering mechanisms to containerise and properly describe processing and analysis steps. Provenance will be adopted to link computational steps with the original data descriptions and their identifiers. Moving computation to the cloud, supported by mechanisms for authentication and accounting and by efficient data staging services is being evaluated in different contexts by a number of ORFEUS-EIDA data centres. A major challenge is to find a sustainable solution that could be adopted and shared by a broader number of data centres.

4.3.2 Technical Activities and Implementation

The FAIR assessments highlighted challenges, extracted from Table 1B, yielded to a set of concrete activities that are summarised below.

- Definition of policies for PIDs. Broader adoption of at least one mechanism to mint PIDs. EPIC Handle seems to be the right candidate for a common solution but customised solutions will be accepted as long as they comply with the agreed policies.
- Population of PIDs in WFCatalog. This metadata catalogue will act as primary source for discovery and access of rich seismic waveform metadata including PIDs.
- Continuation of the work on a common seismology vocabulary and integration of concepts and definitions in the WFCatalog. This task will include work on the data model and the API. Users will be able to request semantic description in JSON-LD format. The descriptions will be compliant with EPOS-DCAT-AP.
- Implementation of a use case addressing computation of seismic waveform data on the cloud supported by the SWIRRL API. In this task the feasibility of provenance collection will be evaluated.

4.4 Task 10.6 - Implementation of FAIR roadmap in satellite Earth Observation community

The task 10.6 is focused on the application of the FAIR principles to satellite Earth Observation products. In particular, the task has to develop awareness and share knowledge about FAIRness in the satellite Earth Observation domain to foster the application of the FAIR principles into existing practices and methods. This task has an effective link with the results of previous and ongoing activities carried out in the context of other projects such as EOSC-hub, EPOS-IP, EPOS-SP and OpenAIRE-Advance. Last but not least, the task has to tackle the issues of the integration with the Copernicus DIAS environments.
The aim is to produce an efficient framework to address the FAIR aspects in the different phases of the product life cycle such as generation, curation, computation, dissemination and publication. The goals of this task are:

- Contribute to the gap analysis carried out in T10.2;
- Harmonise the results of activities in EPOS-IP, EPOS-SP and EOSC-galaxy projects (EOSC-Hub and OpenAIRE-Advance).
- Define and implement technical elements for satellite RIs in order to achieve a FAIRness maturity.

These overarching objectives include the following implementation activities:

- Metadata and quality control (QA/QC)
- Integration and exploitation of DIAS and EOSC computational services
- Collection of requirements for PID mechanism implementation
- Enhancement of AAAI system of Earth Observation RIs.
- Validate and test a set of use cases from Earth Observation community.

4.4.1 FAIR assessment and gap analysis

To perform the FAIR gap analysis, we considered the DDSS provided by the EPOS TCS Satellite Data (SATD) that have been validated and included in the EPOS products portfolio. In particular, we focused on the DDSS referred to as “Line of sight displacement time series”, as reported in table 1A. This DDSS is the most complex and complete among all the services deployed but the TCS SATD and it is affected by all the main FAIR issues of the TCS. Accordingly, this DDSS represents the touchstone of the TCS SATD FAIRness maturity.
The analysed DDSS is released with two components:

- Data, embedded in a csv (ASCII) file, represents a sparse matrix of points located on the Earth surface. For each point a number of fields are provided: a coordinate triplet (latitude, longitude and altitude), a predetermined list of parameters coming from the data processing chain, and a time-series of displacement values. It is worth noting that the time-series does not have a fixed

size, i.e., its length can change dataset by dataset, because the number of samples depends on the number of satellite acquisitions used (processed) to generate the final product;
- Metadata is formatted in a XML file that implements the ISO 19115 standard.

Both data and metadata are stored in the TCS SATD gateway that employs the Geohazards Exploitation Platform, a cloud-based platform developed with the support of European Space Agency (ESA). GEP is an interoperable platform that is queried by EPOS ICS to retrieve data and metadata. The structure of the interoperability layer of the TCS SATD is schematically depicted in Fig. 6
The gap analysis was carried out by critically investigating the several components of the TCS SATD with respect to the requirements of the FAIR principles. Such an analysis provided the gaps reported in table 1A and quickly summarized in table 2 that impact on different FAIR principles.

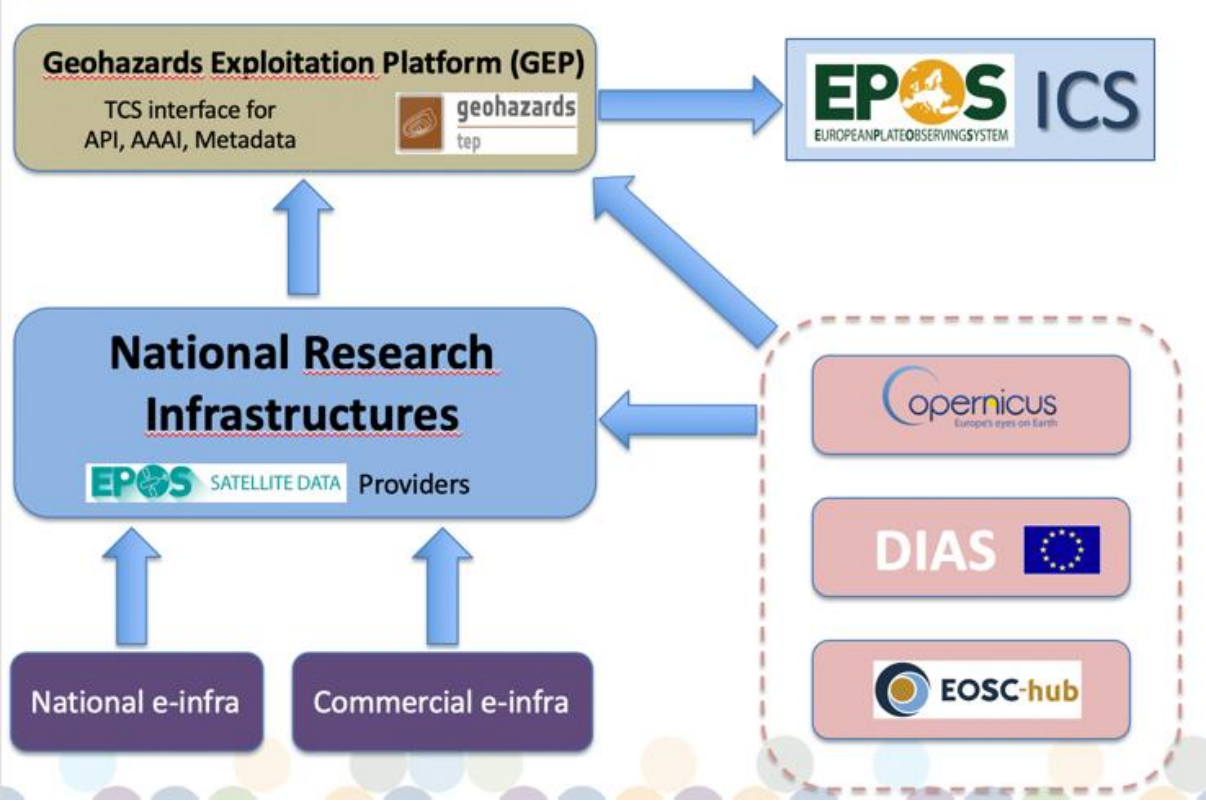

Fig. 6 Schematic view of the structure of the EPOS TCS Satellite Data. The TCS has a unique access point, Geohazards Exploitation Platform, to distribute products and processing services. GEP deals with AAAI, metadata, API, and any other service deployed by the TCS Satellite Data.

Table 2

Gaps	Description	Impact on
PID	No PID system is used and applied	F1, F3 A1
Data and Metadata Life	No policy has been agreed on the lifecycle of data and metadata. Metadata preservation is not guaranteed.	A2
Vocabulary	A preliminary vocabulary has been drafted but the activity is not complete	I2

4.4.2 Technical Activities and Implementation

On the basis of the gap analysis, a list of activities and an implementation roadmap have been drawn up (table 1B).

PID system: this is a long-term activity since it does not have a trivial solution. The adoption of a PID system has a strong impact on the financial and technical sustainability of the RIs. Indeed, a large number of satellite DDSS are live products, because they are regularly updated each time a new satellite image is available (in same cases several acquisitions per week are available). The preservation of all generated products (instead of having a single product regularly updated) could become an unmanageable problem because each product is quite large (up to several gigabytes) and the disk space would exponentially increase. Moreover, even if several PID systems are available (e.g., DOI and handle) the services related to PID management have a cost that has to be supported by the RIs. Accordingly, a PID system suitable for the satellite community has to be found by also investigating its impact on financial sustainability.

Data and Metadata Life: this is a short to medium-term activity; the metadata preservation has no strong impact on sustainability and is technically manageable. It needs an agreement at RIs level to be implemented in the Data Management Plan.

Vocabulary: this is a dynamic action. The building and updating of the vocabulary is an ongoing activity. Once the vocabulary is consolidated, it will be documented and resolvable using globally unique and persistent identifiers.

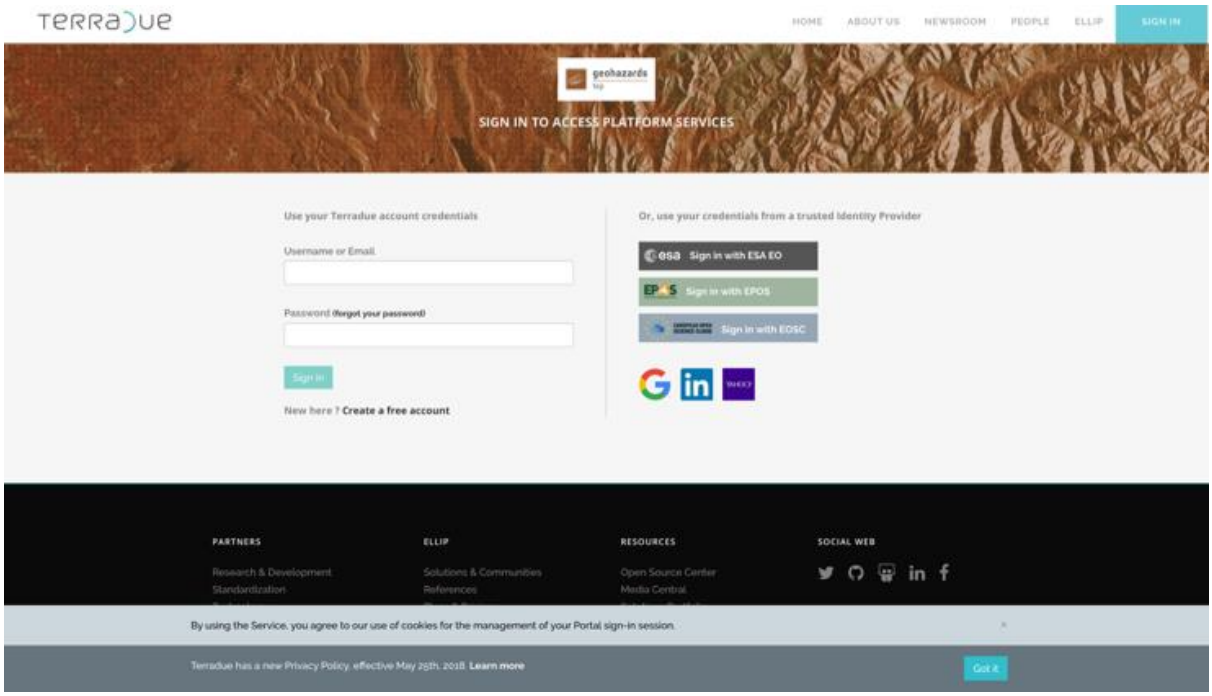

Fig. 7. Login page of the Geohazards Exploitation Platform to access the services offered by the EPOS TCS Satellite Data.

In addition to the above-mentioned activities, the FAIR roadmap of the satellite EO community needs to address the integration with and exploitation of DIAS and EOSC computational services as well as the enhancement of AAAI system of Earth Observation RIs. This latter has been recently already tackled and solved. The TCS SATD has a unique interface where the RIs are integrated; the AAAI system of GEP has been successfully integrated with ICS one and users registered within the EPOS management system can easily access the TCS resources. It is worth noting that the TCS AAAI is already integrated with other systems largely used in the satellite EO community, such as ESA and EOSC (Fig. 7). The integration of satellite RIs with Copernicus DIAS and EOSC computational services represent a more challenging activity. In particular, DIAS, the cloud providers specifically developed by ESA and DG-GROW for Copernicus users, seems to be the more suitable environment where satellite processing services can be effectively implemented. Unfortunately, the efficient exploitation of such an environment needs to properly evaluate the technical and financial sustainability, i.e., the suitability and robustness of the offered services as well as their medium- to long-term competitiveness with respect to similar solutions provided by the commercial sector have to be accurately evaluated. In this framework, the analysis of the DIAS solutions successfully started and it is currently ongoing.

4.5 Task 10.7 - Marine (EMSO)

The prime task objective is to increase the still limited interoperability of geophysical data/metadata of EMSO and EPOS. The adoption of common standards for data and metadata will support scientists in the joint use of land and marine data.
This task will focus on the enrichment of metadata, data and sensors of Ocean Bottom Seismometers (velocity-meter and accelerometers)/ Hydrophones/ Magnetometers (OBS/H/M), portable modules used to set-up networks in marine areas unreachable by land networks or to extend land networks out into marine areas.
Enriched metadata improves data documentation and ensures its re-use. A shared workflow (data curation, long-term preservation), not presently provided across EPOS and EMSO, will be analyzed and agreed for OBS/H/M data to make them findable and accessible over the long term. This task is expected to improve the range of data products and the adoption of FAIR principles across the data pipeline.

4.5.1 FAIR assessment and gap analysis

EMSO ERIC regional facilities do have an essential role in delivering seismic data to broad seismological and geophysical communities, to national agencies and other stakeholders. In fact, because of the still uneven distribution of EMSO marine facilities, a benefit to regional scale seismological and geophysical studies of the Mediterranean and North-East Atlantic shall come only from the regular joint use of EPOS and EMSO data. It is already the case with seismological data at several EMSO regional facilities where standardized seismological data flow to national data centers with links to civil protection and to international seismological agencies. This integration helps to improve the reliability of the localization of the seismicity, especially those events occurring in marine coastal and open sea areas.
The data workflow includes interactions with EPOS through ORFEUS-EIDA. However, the data workflow implemented by the EMSO across regional facilities is not standardized yet. Further, the level of adoption of FAIR principles varies depending on the regional facility. These interactions with EPOS require an extensive standardization of the acquisition and validation process from the sensor level to the recorded metadata, including data format, data transmission protocols, data archiving platforms. Additional service components are based on dedicated methods and software for integrating data from the different observation systems of the facilities and retrieving the basic standard earthquake parameters.
In order to improve the interoperability between EMSO and EPOS, it is crucial for EMSO to develop a harmonization abstraction layer (see Fig. 8). We investigated the process and key features for such a harmonization based on the FAIR principles. A summarized assessment is provided below.

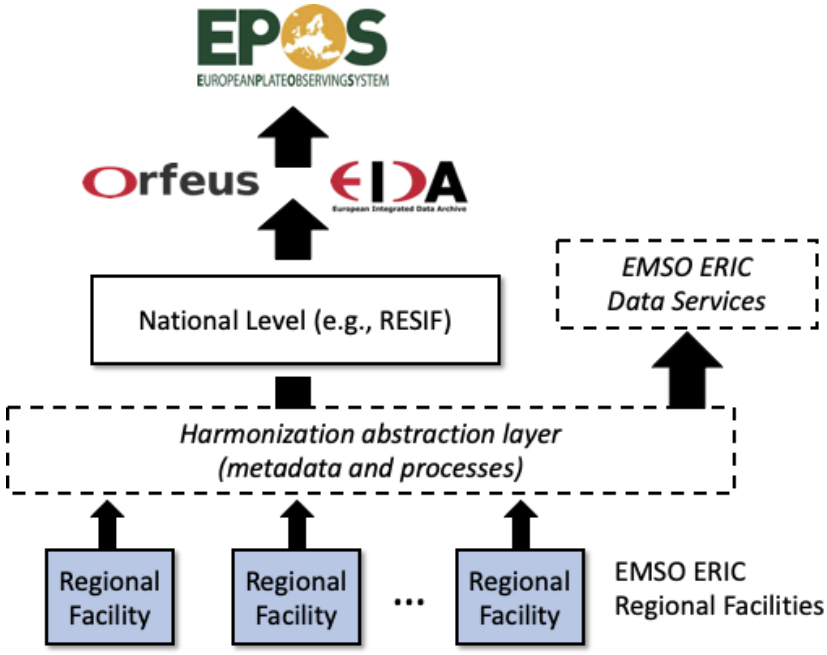

Fig. 8. Seismic data flows from EMSO onward to EPOS. The diagram also shows envisioned extensions (dotted lines).

Findable: While EMSO some data is already findable through standardized mechanisms (e.g., national agencies and ORFEUS-EIDA), not all regional facilities use the same data discovery mechanism. Our goal is to establish the most effective tools to harmonize and enrich metadata to enable discovery and integration into EPOS while providing visibility of EMSO contributions to EPOS.

Accessible: Current access to EMSO seismological data includes different mechanisms such as those from RESIF (French seismological network) and EIDA to distribute data and metadata according to the standards of the International Federation of Digital Seismograph Networks (FDSN), central control systems or specialized applications. The harmonization of mechanisms is essential for accessibility.

Current efforts are focused on better integration with ORFEUS-EIDA and harmonized access interfaces across EMSO regional facilities.

Interoperable: While EPOS represents the front-end in the primary sub-domain, an essential activity within the EMSO back-end is the standardization of processes to ensure interoperability between regional facilities and with other key stakeholders such as EPOS. EMSO will investigate the adoption of standardized seismological vocabularies at the harmonization abstraction layer as they are established.

Reusable: Harmonized and enriched metadata are expected to improve current documentation processes and ensure re-use. A significant challenge is establishing an agreed workflow between EMSO and EPOS (ORFEUS-EIDA), which is necessary to enhance the data curation process.

4.5.2 Technical Activities and Implementation

The implementation of technical activities for the adoption of FAIR principles is preceded by an analysis of available EMSO sources of seismic-related data. In this analysis, we identified:

- regional facilities delivering seismic data;
- types of data produced at each of the regional facilities;
- current processes for providing data and metadata, including ongoing interactions with ORFEUS-EIDA.

Additionally, regional facilities currently not interacting with ORFEUS-EIDA have explored the requirements and interfaces for engaging with ORFEUS-EIDA.

Based on the FAIR assessment and gap analysis, our roadmap for technical activities and implementation includes:

- Implementation of a harmonization abstraction layer (metadata and integration processes).
- Quality control support on seismological data and data product generation.
- Enrich metadata and establish an agreed workflow with ORFEUS-EIDA to enhance the integration and improve the visibility of EMSO contributions through its regional facilities.
- Development of an appropriate FAIRness assessment process, ideally based on both qualitative and quantitative methods (Magagna, 2020).

5 Conclusion

EPOS is a complex environment; it is not a simple portal from which datasets may be downloaded. However, FAIR has been designed-in from the beginning because the architect was involved in defining the Force11 FAIR Principles and since has been involved actively in FAIR discussions through various projects and the RDA FAIR Data Maturity Working Group.

A broad assessment is that EPOS is ‘FAIR enough’. The major principles are covered. However, there are (as always) areas for improvement and these are documented above. The required developments to improve FAIRness for the EPOS ICS-C are fed into the ‘Shape UP’ methodology used in EPOS, prioritised, resourced and executed. The various EPOS ICS-D implementations exhibit a similar state of FAIRness, realised through progressive inclusion of these services in the EPOS ICS-C catalog with rich metadata.

EMSO provides services that are partly FAIR through the EMSO regional agencies. However, there is great benefit in integrating EMSO digital assets with EPOS (via ORFEUS-EIDA) and achievement of this requires a harmonisation layer including provision of rich metadata. Thereafter, FAIRness may be reassessed in this context.

6 References

Bailo, D. et al. (2020). Perspectives on the Implementation of FAIR Principles in Solid Earth Research Infrastructures. Front. Earth Sci. 8, 3.

Blanchard, B. S. (2004). Systems Engineering Management. Hoboken, NJ: John Wiley & Sons, Ltd,.

Magagna, B. et al.(2020) Requirement Analysis Technology Review and Gap Analysis of Environmental RIs. ENVRI-FAIR deliverable 5.1.

Mons, B., Neylon, C., Velterop, J., Dumontier, M., Da Silva Santos, L. O. B., and Wilkinson, M. D. (2017). Cloudy, increasingly FAIR; revisiting the FAIR data guiding principles for the European Open Science Cloud. Inf. Serv. Use 37, 49–56. doi: 10.3233/ISU-170824

Trani, L., Koymans, M., Atkinson, M., Sleeman, R., Filgueira, R., 2017. WFCatalog : A catalogue for seismological waveform data. Comput. Geosci. 106, 101–108. doi:10.1016/j.cageo.2017.06.008

Wilkinson, M. D., Dumontier, M., Sansone, S., da Silva Santos, L. O. B., Prieto, M., Batista, D., et al. (2019). Evaluating FAIR maturity through a scalable, automated, community-governed framework. Sci. Data 6:174. doi: 10.1038/s41597-019-0184-5

7 Annex A - Glossary

AAAI : Authentication, Authorisation, and Accounting Infrastructure
CERIF: Common European Research Information Format
DCAT: Data Catalog Vocabulary
EPOS-DCAT-AP: DCAT Application Profile for EPOS
DDSS : Data, Data products, Software and Services
DIAS : Data and Information Access Services
EOSC : European Open Science Cloud
EPOS Strategic Plan : Defines the activities of EPOS-ERIC
GEP : Geohazards Exploitation Platform
ICS-C : Integrated Core Services - Central Hub
ICS-D : Distributed Integrated Core Services Distributed
PID : Persistent Identifier
SDLC: Software Development Life Cycle
TCS : Thematic Core Services