D10.1
TECHNICAL ANALYSIS AND DEFINITION OF IMPLEMENTATION COMPONENTS FOR FAIR IMPLEMENTATION OF RIS IN THE SOLID EARTH SUBDOMAIN

Work Package	WP10
Lead partner	INGV
Status	Final
Deliverable type	Report
Dissemination level	Public
Due date	31/12/2019
Submission date	30/12/2019

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

Deliverable abstract

D10.1 describes the state of readiness for FAIR and EOSC achieved by EPOS, the RI for the solid earth sub-domain. With a questionnaire, the state of EPOS assets was collected and analysed. This was supplemented by a separate small survey using the CSIRO method. The current state of implementation of EPOS is described and also the plans for the near-future. Additionally, the transitional governance and management arrangements for EPOS – which control the overall policy and hence FAIRness – as we move from EC-funded implementation to EPOS-ERIC managed operational state is described.

The results of the main questionnaire activity indicated that EPOS may be described as FAIR while the survey using the CSIRO method indicated some shortcomings. FAIR is not a state but a journey and there is always room for improvement.

The results concerning readiness for EOSC indicate that – because the EPOS catalogue is based dominantly on services (at present: datasets, workflows and other assets are being worked upon) it is compatible with the architectural direction of EOSC. However, analysis of the current proposed metadata schema for the EOSC services catalogue indicates a need for discussions to ensure it is sufficiently rich for the purposes of EPOS.

The conclusion is that EPOS is both FAIR and EOSC-ready. This is mainly due to the design of the EPOS architecture which took into account from the beginning the need for FAIRness and the direction of development (i.e. towards a service catalogue not a data catalogue) of EOSC.

DELIVERY SLIP

	Name	Partner Organization	Date
Main Author	Riccardo Rabissoni	INGV	29/11/2019
Contributing Authors	Keith Jeffery Jean-Baptiste Roquencourt Daniele Bailo Sylvain Grellet Abdelfettah Feliachi	UKRI-BGS BRGM INGV BRGM BRGM	29/11/2019
Reviewer(s)	Florian Haslinger	ETH Zürich	12/12/2019
Approver	Andreas Petzold	FZJ	30/12/2019

DELIVERY LOG

Issue	Date	Comment	Author
V 0.1	2019-11-19	Version for WP leader and Internal Review	Keith Jeffery
V 0.2	2019-11-29	Revised V0.1 with input from BRGM	Daniele Bailo
V 0.3	2019-12-12	Revised after comments from WP10 team	Keith Jeffery

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

Introduction

The ENVRI-FAIR project is engaging Research Infrastructures (RIs) in the environmental domain covering the subdomains Atmosphere, Marine, Solid Earth and Biodiversity / Ecosystems. The overarching goal of ENVRI-FAIR is that all participating Research Infrastructures (RIs) will improve their FAIRness and become ready for connection to the European Open Science Cloud (EOSC). WP10 has a focus on the Solid Earth subdomain, represented by the European Plate Observing System (EPOS) RI [1] – a landmark in the ESFRI roadmap, now an ERIC - which engages 10 different scientific communities - TCS: Thematic Core Services) with the goal of merging metadata descriptions of their assets into a single, centralized FAIR hub, namely the Integrated Core Services (ICS), for accessing sub-domain (community) specific heterogeneous assets. External common e-Infrastructures or specialised community services are grouped as ICS-D (ICS Distributed) (Figure 1). EPOS is the single consolidated RI related to this sub-domain.

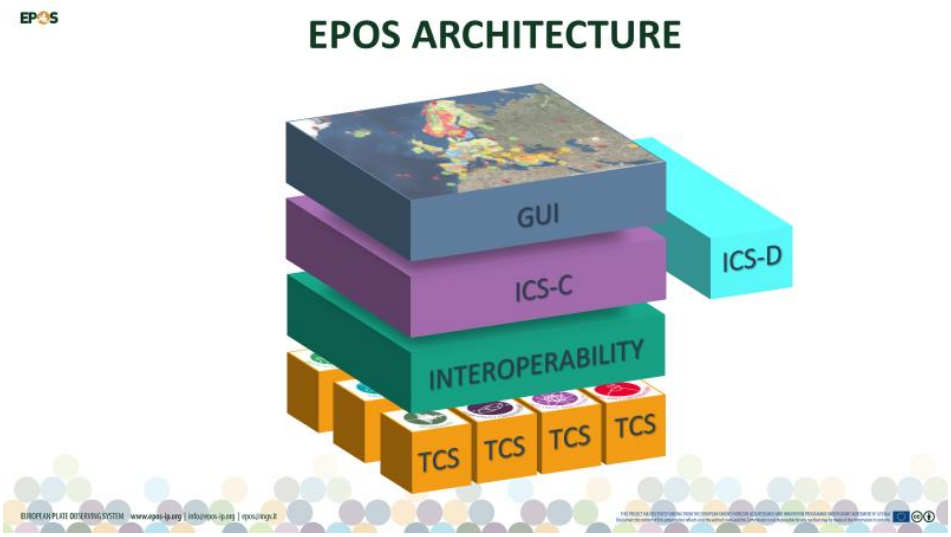


Figure 1: EPOS Architecture.

The final scope of WP10 is first to assess the FAIRness of EPOS by analysing the maturity status of the scientific communities that are part of its ecosystem, in order to implement the necessary actions to improve the adherence to the principles. The whole process is carried on in close and regular interaction with WP5-WP7 that consider common standards, training, common implementation options at environmental domain level, as well as with WP8-WP11, which concern analysis and implementation activities in each of the subdomains.

This report is the result of Task 10.1 "Harmonised analysis of FAIR principles at subdomain level" and partially follows the outcomes from Task 10.2 "Harmonised definition and harmonised implementation of FAIR principles at subdomain level".

Background

EPOS for solid earth domain

The aim of EPOS is to provide a single RI (Research Infrastructure) for the earth science sub-domain. A key component – alongside governance, harmonisation, financial sustainability – is the leading-edge ICT required to achieve this aim.

TCS

Within EPOS 10 communities have coalesced from multiple international and national initiatives and institutions to provide TCS: Thematic Core Services. These services are software services that provide access to, manipulation of and other operations (such as analytics, visualisation) on geoscience data. Further communities are in the process of joining.

ICS-C

ICS-C (Integrated Core Services – Central) is the portal for finding, accessing, interoperating and reusing the data assets of EPOS through the services provided by the TCS. ICS-C provides:

1. Services to ingest metadata describing the digital assets from the TCS, converting from the 1 different metadata formats used locally within the communities to a single canonical rich metadata format CERIF (Common European Research Information Format: an EU Recommendation to Member States) [2]; the pragmatic approach was to leverage EPOS-DCAT-AP [6] within the ingestion pipeline.
2. Services to discover digital assets by querying the rich metadata catalogue constructed by the services in (1);
3. Services to display the output of each of the selected TCS services by invoking them from ICS-C;
4. Services to store the metadata for relevant digital assets in a workspace ready for use in a workflow;
5. Authentication Services for AAAI.

Current developments extend the capability of ICS-C as follows:

1. Assisting in constructing a workflow from the contents of the workspace;
2. Deployment of workflows to future ICS-D (Integrated Core Services Distributed);
3. Authorisation services for AAAI (Authentication, Authorisation, Accounting Infrastructure);
4. Provision of a TNA (Trans-National Access) service to support access to laboratory or sensor networks of one organisation from one or more users at another;

Thus ICS-C provides facilities for end-users to utilise:

1. Discovery – finding;
2. Contextualisation – accessing;
3. Interoperating using harmonised metadata from converters for 17 metadata formats to one canonical rich metadata standard;
4. Reusing – for different communities e.g. near fault observatories using seismic services;

trusted through an agreed validation process - the DDSS (Data, Data products, Software, Services) provided by the community are validated by the community itself and EPOS SCB (Service Coordination Board which under EPOS-ERIC transforms to SCC (Service Coordination Committee)) where leaders of the TCS

communities together with IT team representatives make recommendations to the EPOS Executive Director. They are also monitored to demonstrate their sustainability as well as serving only available validated DDSS. EPOS thus provides FAIR assets for geoscience.

CERIF

EPOS adopted the use of a canonical rich metadata format. This means instead of converting each metadata format to every other one (brokering) which would require n(n-1) convertors, EPOS requires only n convertors. This allows each existing relevant metadata format to be represented in a homogeneous manner after conversion thus providing (a) one homogeneous representation of all the metadata to support interoperation since every metadata representation of a digital asset has the same syntax and semantics; (b) a facility for converting from one metadata format to another if required for particular purposes. There is a balance between what can be provided by the TCS communities and what is required in the catalogue for optimal operation of the ICS.

CERIF is the only known metadata format to represent a fully-connected graph using n-tuples. As well as both formal syntax and declared semantics, this also allows assurance of referential and functional integrity. CERIF in its syntactic layer represents things in the real world as base entities (objects) e.g. person, dataset. It represents relationships between them as linking entities (objects) which relate together the two referenced/linked base entities with a role (e.g. owner) and a temporal duration. Records in entities are n-tuples and so richer than the triples of RDF (Resource Description Framework). This structure supports both curation and provenance concepts. The semantic layer has the same structure as the syntactic, with base entities being vocabularies and linking entities being the relationships between terms. This provides full ontological capabilities and permits multilinguality while ensuring consistency by unique instantiation. Thus roles, or coded attribute values, in the syntactic layer are pointers to terms in the semantic layer.

CERIF is also the underlying data model for OpenAIRE and provides the basis for ORCID. As requested by the European Commission, CERIF is maintained, developed and its use promoted by the euroCRIS organisation.

DCAT

DCAT [3] is the dataset catalogue vocabulary defined by the W3C (World-Wide Web Consortium) to represent metadata about datasets using RDF. It is a model that provides to the publishers a way to expose catalogue components: catalogue records, datasets, distributions, etc., and the relations between them. One of the main goals of DCAT is to enable decentralized catalogue description and federated search across multiple sources. The Joinup European platform proposes DCAT-AP (Application Profile) [4] as a specification of this model. One of the main purposes of this model extension is to provide content aggregators to gather catalogue descriptions in common portals such as the European Data Portal [5].

In the framework of EPOS, a specific DCAT Application Profile was defined, EPOS-DCAT-AP [6], in order to facilitate metadata ingestion from asset owners/suppliers (with many heterogeneous metadata formats) through EPOS-DCAT-AP to the EPOS ICS-C CERIF catalogue. GeoDCAT-AP [7] has been proposed as part of the INSPIRE Platform to describe metadata elements for geospatial data. However, APs preclude full interoperability since only the 'core' is common across application profiles and AP vocabularies are domain-specific.

A second version of DCAT (DCAT-2) [8] has been proposed lately (nov-2019) as a W3C recommendation. This second version overcomes some of the inadequacies of DCAT and provides a revision to the DCAT model to include the requirements that emerged from the DCAT use cases and experience. This version can catalogue services in the same way as datasets. It also provides a way to represent more details about the provenance and the quality of datasets.

FAIRness

EPOS considered what is now termed FAIRness from the beginning and EPOS staff were involved in the definition of the FAIR principles and have since been involved in appropriate RDA Working Groups and other discussion fora. FAIR is intended to characterise datasets. The problem with providing FAIR for data was anticipated by EPOS, namely the problem of computing resources to handle the data FAIRly and the network latency involved in simple download of data to a user-defined location (usually their own computer). Furthermore, the EOSC (European Open Science Cloud) is based on the concept of a catalogue of services. By adopting a catalogue of services EPOS has anticipated this approach. As defined in the Force 11 FAIR Principles foundation document [9], FAIR is for machines as well as people. Thus, to assess EPOS FAIRness one has to take into account how EPOS is FAIR with regards to those two really different users.

FAIRness in the framework of EPOS

The process of "making EPOS more FAIR", also known as the FAIRification process, includes several steps, the first one being the FAIRness assessment. After that, a clear gap analysis leads to the implementational activities needed to make FAIR a reality. In the context of EPOS, it emerged that FAIR principles are not readily or practically understandable by RI implementers and data practitioners, so they were reorganized according to an approach that is better understood by domain scientists, the so called four stages roadmap described in 3.4. This may serve as a paradigm to focus, from an RI perspective, on the actual FAIR requirements and consequent activities needed to comply with FAIR principles.

FAIRness assessment

Assessing FAIRness of such an infrastructure in the framework of ENVRI-FAIR is hard in one pass as:

- EPOS as a whole is made up of sub-communities and their corresponding IT approaches (often siloed);
- The EPOS ICS-C catalogue concerns the description of datasets, services, people and organisations, facilities, equipment and other entities exposed by each asset owner/supplier converting local metadata formats to EPOS-DCAT-AP as the intermediate format for onward conversion to CERIF [10] for the ICS-C catalogue.

There are different initiatives aiming at providing a framework to self-assess the level of FAIRness of a given research infrastructure datasets and services; for instance, the RDA FAIR data maturity Working Group [11] provides the forum to discuss those frameworks.

Two approaches have been used here in order to assess EPOS level of FAIRness.

- Implementing the GO-FAIR based questionnaire [12] and incorporating/reusing a pre-existing self-assessment framework stemming from RDA FAIR data maturity Working Group;
- Applying the CSIRO assessment method [13].

FAIRness evaluation in ENVRI-FAIR

FAIRness assessment was done in synergy with WP5. The questionnaire is partially inspired by the one proposed by the GO FAIR initiative [11].
The FAIR questionnaire submitted to the EPOS community is a set of fewer than 80 questions to assess the use of standard vocabularies, repository maintainability, easiness of service findability, potential plans for data management. All of the possible answers have been intentionally presented as open text, to encourage people to provide as many details as they wished.
Then, a total amount of 8 significant contributions from different subdomains of the Solid Earth ecosystem - Seismology, Near-Fault Observatories, GNSS Data and Products, Volcano Observations, Satellite Data, Geological Information and Modelling - have been collected and a work of harmonisation among the relative answers has been carried out. The questionnaire is available in Annex A - Questionnaire.

Another potential evaluation tool is from the CSIRO OzNome initiative [13], seeking to connect information infrastructures across Australia and enable researchers, industry and key partners to achieve productivity gains around their discovery, access and use of data.
In a limited trial, this self-assessment rating tool has been applied to EPOS as follows:

- The EPOS-DCAT-AP representation of the CERIF catalogue at EPOS ICS-C concerning the description of datasets, services, people organisations, facilities, equipment etc. exposed by each asset owner/supplier converted from local metadata format to EPOS-DCAT-AP and thence to CERIF;
- A selected 'typical' dataset or service exposed by TCS: that was the only way deemed realistic compared to applying it to the whole breadth of EPOS TCSs datasets and service.

It was done in two parallel CSIRO (OzNome) 5-stars rating tool exercises to avoid mixing both ICS-C and TCS concerns in the replies (thus the evaluation). Indeed, this proved to be a limiting element when answering the questionnaire as people interviewed were often both active in ICS-C and some TCS
Results from the first one are reported in Figure 2 (a), results from the second one in Figure 2 (b).

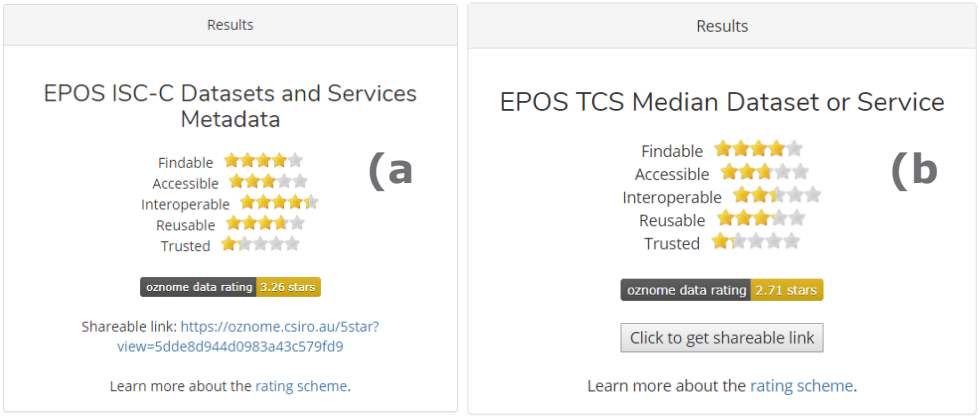


Figure 2: Results from CSIRO Questionnaire.

Detailed answers to the questionnaire (a) are available in Annex B – OzNome 5-stars questionnaire answer, or online following this link:
https://oznome.csiro.au/5star?view=5dde8d944d0983a43c579fd9

The detailed answers to the questionnaire (b) are available in Annex B – OzNome 5-stars questionnaire answer, or online following this link: https://oznome.csiro.au/5star?view=5dde956b4d0983dfe2579fda

The CSIRO method is very much biased towards the use of LOD (Linked Open Data), RDF and OGC standards. Since EPOS is using a different (and richer) technology unsurprisingly the example assets analysed did not score highly.

Gap analysis

The gap analysis was based on the questionnaire survey. All the answers have been human-interpreted to schematize and harmonize the common ones. Hence, every input has been transformed into a yaml file (example in Annex C – yaml files), and subsequently ingested via a Jupyter Notebook script. Afterwards it has been SPARQL-queried to perform all the necessary analysis on the collected data (Annex D – SPARQL example).
This latter activity permits the usage of questionnaires as a benchmark for performing a semi-automated evaluation, thus providing measurable criteria for gap analysis. At the moment, unlike the possibilities given by tools like OzNome, the method followed in the questionnaire provision – with its subsequent analysis – is not so quantitative since it doesn't provide a specific score of how FAIR is the observed RI. Its principal strength lies in the fact that it allows, by performing appropriate queries on the collected data, to have a glance (and concentrate) on the aspects that have been implemented less (or at all) with respect of a specific principle of the F-A-I-R.

A pyramid for contextualization of principles

Once the FAIRness of each Research Infrastructure of the four subdomains is evaluated and assessed, it comes to actual activities for filling in the gaps. This requires a clear approach, with a defined process to handle the activities needed to move from principles to FAIRness assessment to actual implementation.

In the Solid Earth sub-domain, leveraging on years of experience with the actual thematic communities, we realized that domain scientists, practitioners and managers have a common approach, which is reflected in the re-organization of the detailed FAIR principles [9] into a four-stages roadmap (see Figure 3) which include: a) data stage, b) metadata stage, c) access stage and d) (re-)use stage. These stages correspond to the actual conceptual approach driving day-to-day work of RI implementers in the solid Earth domain (EPOS).

Data are usually the main business and wealth of scientists and data practitioners in RIs. As a consequence, the first conceptual step relates to data aspects (Stage 1) and has as direct consequence the challenge of data description by means of metadata and identification that, in order to create the premises for data searchability and contextualization (Stage 2), needs also to be tackled. The other challenge is providing access to data by means of appropriate technologies (Stage 3). In order to include functionalities that go beyond data access, for instance data analysis and processing, FAIR RIs and data stewardship systems should address a fourth stage concerned with services that make use of data (Stage 1) and metadata (Stage 2) FAIRly accessed (Stage 3) and produce new (meta)data as output.

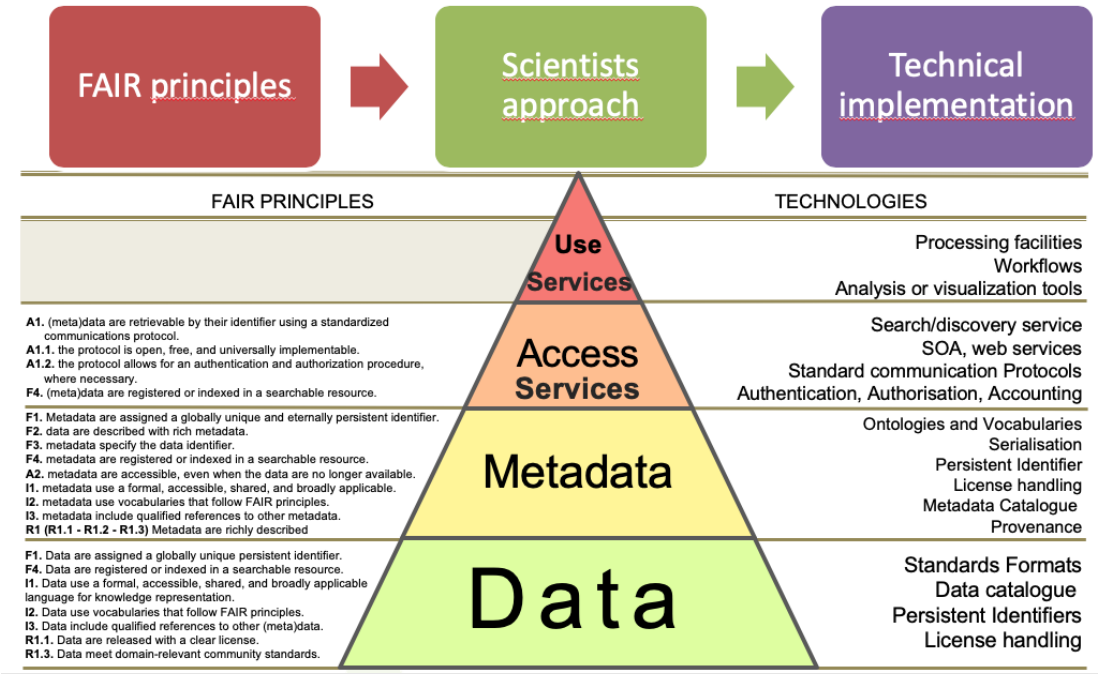


Figure 3: The pyramid represents the approach of data practitioners, scientists and managers within the thematic communities for developing FAIR services that provide access to FAIR datasets. On the left, detailed FAIR principles are recognized according to the pyramid layer. On the right, technologies at each of the layers are suggested in order to fulfil a specific FAIR detailed principle.

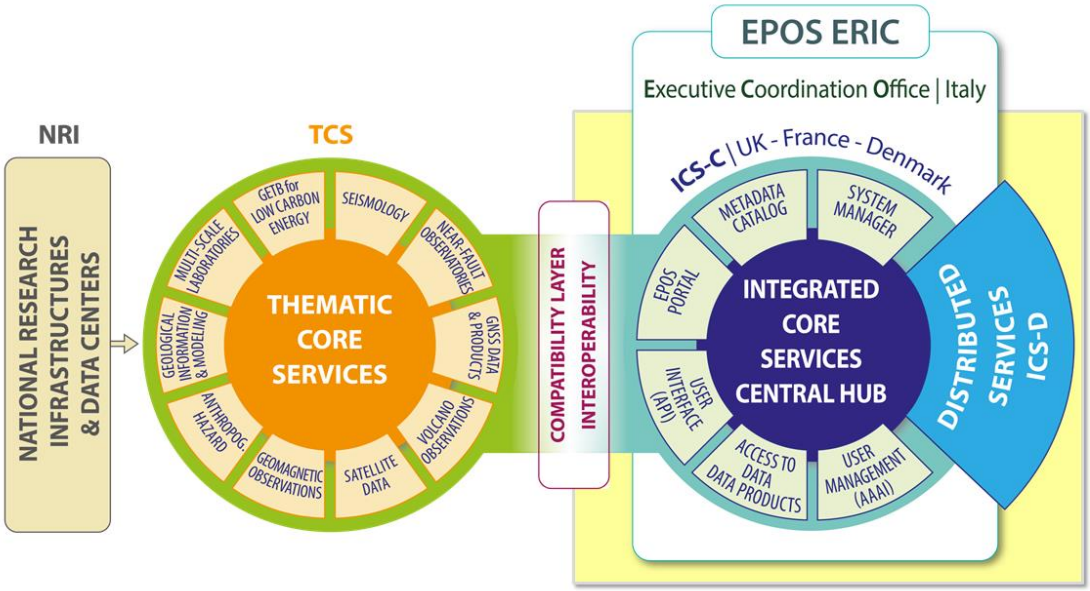
After reorganizing FAIR principles and relating them to stages that correspond to the mindset and approach of the scientist and data practitioners, IT professionals and managers, technologies were selected in order to implement the FAIR requirements evidenced at each stage of the pyramid. Such technologies do NOT refer to specific software packages, programming languages or other technical detailed approaches. They are rather reference technologies that are likely to be needed to meet any of the requirements of the corresponding FAIR detailed principles.
For instance, if an IT developer wants to fulfil F2 (data are described with rich metadata) at the metadata layer, then it is likely that a metadata catalogue needs to be implemented, with all related work (e.g. creating or re-using a canonical ontology for representing metadata).

The method described above for assessment of FAIRness and pointing to the requirements to achieve FAIRness in a RI is comprehensive and leads to an implementation roadmap.

Implementation components

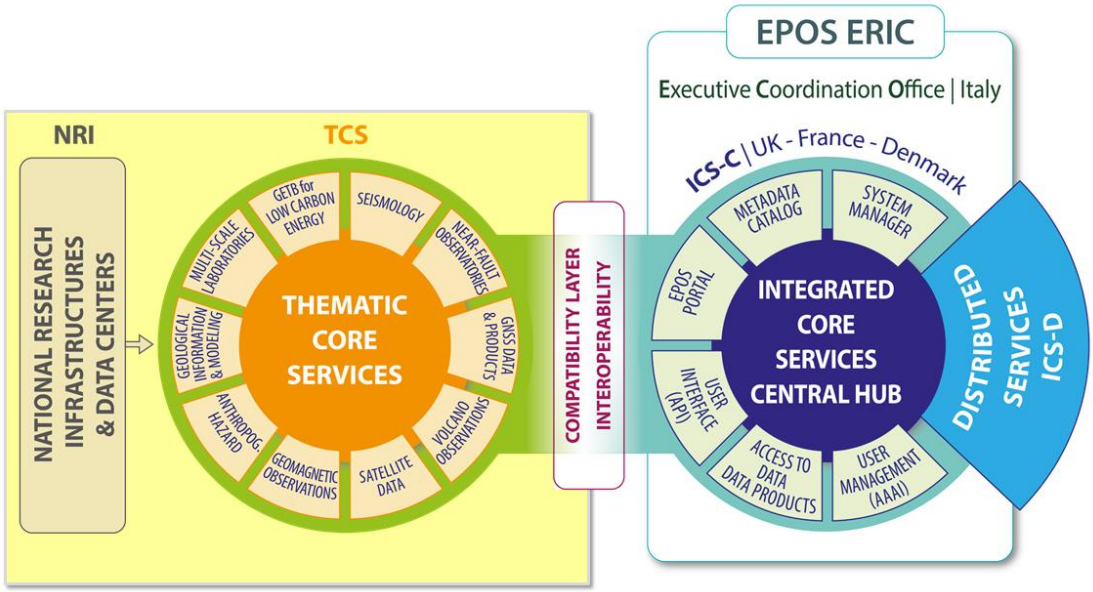
EPOS architecture

As indicated in Section 2, the EPOS ICT architecture is based initially on services providing access to data assets (and other assets) at the multiple RIs within each TCS (the 'treasure' of EPOS) as highlighted in yellow in Figure 4 being described by harmonized rich metadata at the ICS as highlighted in yellow in Figure 5, thus providing homogeneous FAIR utilization of the heterogeneous assets. Other assets (e.g. datasets) are being added progressively.



Main elements of the EPOS Architecture, the Integrated Core Services Central Hubs (ICS-C) and the Executive and Coordination Office (ECO) belong to the EPOS-ERIC legal subject.

Figure 4: The TCS part of the EPOS Architecture.



Main elements of the EPOS Architecture, the Integrated Core Services Central Hubs (ICS-C) and the Executive and Coordination Office (ECO) belong to the EPOS-ERIC legal subject.

Figure 5: THE ICS component of the EPOS Architecture.

The canonical rich metadata catalogue uses CERIF as described in Section 2. This provides the Findability, Accessibility, Interoperability and Reusability required. The ICS-D (Integrated Core Services - Distributed) component of the ICS part of the architecture will allow workflows of services to be executed utilizing other e-Infrastructures (such as supercomputers or sensor networks) within or independent of any research infrastructure in the TCS communities.

In this architecture, EPOS-DCAT-AP has been used since year 2 of EPOS-IP between the multiple metadata formats of asset owners/suppliers in the TCS and the ICS-C CERIF catalogue. This mediation layer enables a homogeneous description of the heterogeneous assets exposed by each TCS. In addition to

using this syntax to describe their assets, TCS IT teams also use a common set of keywords (semantics) defined progressively by 'harmonisation teams' across the TCS.

The EPOS ICS-C Web API (Swagger/OpenAPI) enables external components & systems to search the EPOS ICS-C CERIF catalogue. This approach also enables the contents of the EPOS CERIF catalogue to be represented in many other metadata formats to increase interoperability with systems using those other formats. An example is using GeoJSON (Figure 6).

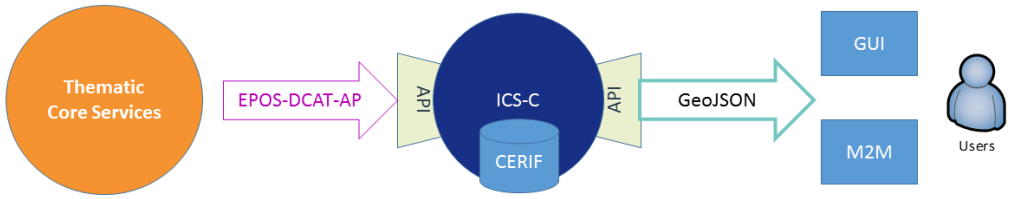


Figure 6: EPOS-DCAT-AP and CERIF in EPOS ICS-C.

Current state of implementation

Organisation and Governance

EPOS is operating in pre-operational mode from October 2019 to end-September 2020. The EPOS-ICS system runs smoothly with a catalogue describing ~200 digital assets in the TCS communities. EPOS is – under the ERIC mechanism rather than as an EC-funded project – leveraging the lessons learned from EPOS-IP. The project is foreseeing the future IT governance with three teams: (a) operation; (b) development; (c) user requirements and ICS-TCS interaction.

These teams are achieved respectively by:
(a) a consortium of three organisations: BGS: British Geological Survey; BRGM: Bureau de Recherches Géologiques et Minières; GEUS: GEologiske UnderSøgelse, supplying an operational e-Infrastructure mirrored across two organisations and supported (user issues, training, monitoring) by a third. For sustainability purposes, these three organisations are engaged with EPOS-ERIC in a Collaboration Agreement;
(b) a group of organisations led by INGV (Istituto Nazionale di Geofisica e Vulcanologia) and formed from the former WP7 from EPOS-IP H2020 project that has a list of issues to be fixed (resulting from ICS-TCS workshops and User Feedback Group meetings) and new technical capabilities to be implemented continuously under EPOS-ERIC;
(c) a group of organisations led by UiB (University of Bergen) that worked with the TCS communities to acquire digital assets, describe them with metadata, assist in metadata conversion and record the assets for costing and governance purposes and also responsible for collecting and evaluating new requirements from the TCS communities. This continues under EPOS-ERIC.

It is intended to maintain the manpower and knowledge from EPOS-IP WP6-7. They will be combined with the experience from the three organisations providing the operational e-Infrastructure to move from a prototype to a product.

A senior technical manager will lead the three teams, and she will find assistance in a committee composed of leaders from the three teams, will report to the EPOS-ERIC Executive Director and be governanced by the SCC, taking into account any advice from the external advisory board.

Technical

The EPOS IT system has been developed in stages. First, as a feasibility prototype in the EPOS-PP (Preparatory Phase) project that demonstrated the flexibility and robustness of the architecture for user-defined requirements, confirming feasibility. Second, as a user-tested prototype in the EPOS-IP (Implementation Phase) project that demonstrated the capability to support a range of user requirements of varying complexity and capability to sustain some load of concurrent users.

During the transition period to September 2020 the operational system will be ruggedized with associated further testing by TCS communities to assure that flexibility, resilience and performance are fit-for-purpose.

The current system is implemented using an agile, spiral development method. It is intended – at an appropriate time - to consider a move to DEVSECOPS [14]. A pipeline to assure development governance with quality control has been adopted. The architecture is based on a micro-services approach which are atomically functional with clean interfaces allowing replacement should a superior module become available. Internally, the components intercommunicate via a software bus using standard APIs.

The catalogue is maintained within a relational Database Management System (Postgres) for performance, integrity, reliability and flexibility. The CERIF complex records in the catalogue can be converted to other metadata formats and in particular to a linked data structure using RDF (Resource Description Framework) to allow interoperation with systems choosing this style of metadata representation.

Implementation plan

The main elements of the plan yet to be completed are as follows:

1) Generic integration of ICS-D: at present pilot projects have been and are being pursued to understand the required metadata elements to describe the workflows and e-Infrastructures to allow automated assistance in workflow construction;
2) Further generic integration of ICS-D: the provision of middleware for optimized deployment of a validated workflow across multiple e-Infrastructures based on parameters of cost, performance, data locality and appropriate right and permissions;
3) Generic mechanisms for authorization within AAAI: this involves additional metadata elements concerned with each digital asset described in the catalogue and appropriate access control software at the ICS-C to authorize access to services in the TCS communities and ICS-D;
4) TNA (Trans-National Access): a prototype service to manage TNA has been implemented which makes available information on appropriate assets hosted at specific institutions (e.g. laboratory equipment, detector networks) to which access may be applied for by researchers from other organizations. Currently there are discussions over governance, financing and the 'rules of access' and once these are concluded the implementation to operational status will proceed;
5) Provision of convertors to/from CERIF as necessary to ensure maximal openness and FAIRness with regard to systems using metadata standards other than CERIF;

6) Improve the level of coherence of the metadata and data formats for the different TCS to improve their interoperation and reuse;
7) Other Solid Earth subdomain developments to ensure greater FAIR compliance within their services.

Conclusions and next steps

EPOS claims already to be FAIR. This has been achieved by concentrating on services to access data where the services are owned, managed and described by the service provider who also provides the required computing resource for execution of the service.
The description of the datasets and services using a canonical rich metadata format is one key to FAIRness providing homogeneous access over heterogeneous assets, together with the definition and implementation of an EPOS-DCAT-AP vocabulary for collecting TCS metadata.
However, FAIRness can always be improved. From the FAIRness evaluations carried out in this task and reported here, various enhancements are identified and remain to be considered for possible implementation. This would be used as a basis for the planned D10.2 deliverable ('Roadmap for implementation of FAIR concepts') where actual tasks, technologies and building blocks will be defined. Thus, from its current state of implementation and implementation plan, EPOS can be considered both FAIR and EOSC-ready.

References

[1] www.epos-eu.org
[2] https://www.eurocris.org/cerif/main-features-cerif
[3] https://www.w3.org/TR/2014/REC-vocab-dcat-20140116/
[4] https://joinup.ec.europa.eu/solution/dcat-application-profile-data-portals-europe/release/200
[5] https://www.europeandataportal.eu/
[6] https://github.com/epos-eu/EPOS-DCAT-AP
[7] https://inspire.ec.europa.eu/good-practice/geodcat-ap
[8] https://www.w3.org/TR/vocab-dcat-2/
[9] https://www.force11.org/group/fairgroup/fairprinciples
[10] https://www.eurocris.org/cerif/main-features-cerif
[11] https://www.rd-alliance.org/groups/fair-data-maturity-model-wg
[12] https://www.go-fair.org/
[13] https://publications.csiro.au/rpr/pub?pid=csiro:EP175062
[14] https://www.devsecops.org/blog/2015/2/15/what-is-devsecops

Annex A - Questionnaire

Nr	Question
1	Date of response
2	Version
General	
3	Contact name
4	Email
5	Research Infrastructure acronym
6	Research Infrastructure Name
7	Research Infrastructure Website
8	Please indicate in which domain your RI is mainly working
9	Please provide the URL of one of the datasets in scope for your answers
10	Please provide the URL to the discovery portal in which the dataset can be downloaded
Repositories	
11	Please provide the URL of the repository you use
12	Please provide the name of the repository
13	Which kind of repository is this?
14	How is the repository within your Research Infrastructure organised?
15	Which repository software is being used?
16	Do you use persistent identifiers or local IDs?
17	If you use PID's, which PID system do you use?
18	Do your identifiers resolve to a landing page?
19	Do you assign identifier manually or automatically?
20	Which identifier registration provider do you use?
21	Is the identifier described with metadata? According to which schema?
22	Is the repository certified? If so, which methods are used?
23	Are repository policies mentioned at the website? If so, indicate the major ones.
24	Are your repositories registered in a registry? If so which registry?
25	Which persistency guaranties are typically given?
Access mechanisms	
26	How is authentication done?
27	Please provide a URL to the description of the Access Protocol
28	Does the protocol allow open access?
29	Do you maintain an own user database?
30	Do you use a person identification system in your AAI? Which one?
31	What is the major access technology supported?
32	How is authorization done?
33	Authorization is required to access the content of my RESOURCE ID
34	Which specific licenses do you use for your data?
35	Please provide the IRI for your usage license regarding the content returned from RESOURCE ID (be that data, or metadata):
36	Are metadata openly available?
Data	
37	Which are the most popular data types used?
38	Which are the preferred data formats?
39	Do those formats include metadata headers? if so, which?
40	Did you register your schemas in a common registry?
41	Do you provide search on data?
Metadata	
42	Please provide the URL of the metadata schema used
43	What is the name of the metadata schema?

44	How is provenance included?
45	Do you provide machine readable provenance information about your data (based on PROV-O or similar)?
46	Are all categories used in the schemas defined in open registries?
47	Are PIDs included in the metadata description?
48	What is the primary storage format for metadata?
49	Which are the export formats supported?
50	Are your metadata made available for search engine indexing?
51	Which metadata exchange/harvesting methods are supported?
52	Do you have a local search engine?
53	Do you support external search engines?
54	Do you make statements about access policies in your metadata?
55	Please provide the URL to a metadata longevity plan
56	Is your metadata machine actionable?
57	Please provide the IRI to a document that contains machine-readable metadata for the digital resource
Semantics	
58	Please provide the URL of the semantic vocabulary in use
59	Indicate the vocabulary name
60	What type of vocabulary is it (taxonomy, thesaurus, ontology)?
61	Indicate the vocabulary topic (generic, domain-specific, project-specific)
62	Which vocabulary language is used?
Data Management Plans	
63	Do you use or provide specific DMP tools? If so, which DMP tool are you using or advocating in your community?
64	Do you apply special data publishing steps?
65	Do you use a community compliance validation service for data?
Data processing	
66	Do you apply special data [processing] steps?
67	Do you apply workflow frameworks for processing your data?
68	Do you use distributed workflow tools? if so, which?
69	Do you offer other type of support or analytics services?
70	Do you offer data products in your RI?
FAIRness	
71	Do you believe that your data is Findable (F)?
72	Indicate where you see major gaps.
73	Do you believe that your data is Accessible (A)?
74	Indicate where you see major gaps.
75	Do you believe that your data is interoperable (I)?
76	Indicate where you see major gaps.
77	Do you believe that your data is re-usable (R)?
78	Indicate where you see major gaps.

Annex B – OzNome 5-stars questionnaire answer

The detailed answers to the questionnaire about the ICS-C Datasets and Services Metadata are:

Questionaire

Tell us about your data

... publication and indexing

1. * Dataset identity

Dataset name or title	EPOS ISC-C Datasets and Services Metadata
URL	http://ics-c.epos-ip.org

2. * Published - is the data accessible to users other than the creator or owner?

- ○ No
- ○ By individual arrangement
- ○ File download
- ○ Institutional or community repository
- ○ Bespoke web service (informal API)
- ◉ Bespoke web service (OpenAPI/Swagger)
- ○ Standard web service API (e.g. OGC)

3. Citeable - denoted using a formal identifier

- ○ Not citeable
- ○ Local identifier
- ◉ Web address (URL - not guaranteed stable)
- ○ Persistent web identifier (URI)

4. Described - tagged with metadata

- ○ No metadata
- ○ Abstract and keywords
- ○ Basic metadata (e.g. Dublin Core)
- ○ Specialized metadata (e.g. Darwin Core, ISO 19115/19139, schema.org scientific data profile)
- ◉ Rich metadata using multiple standard RDF vocabularies (e.g. DCAT, PROV, ADMS, GeoDCAT, FOAF, ORG, GeoSPARQL)

5. Findable - indexed in a discovery system

- ○ no
- ○ local or internal system only
- ◉ community wide or jurisdictional system
- ○ highly ranked in general purpose index (Google, Bing etc)

Tell us about your data

... linked and useable

6. Loadable - represented using a common or community-endorsed (i.e. standard) format

- ○ bespoke format (text, binary)
- ◉ one standard format, denoted by a MIME-type
- ○ multiple standard formats

7. Useable - structured using a discoverable, community-endorsed (standard?) schema or data model

- ○ no formal schema
- ○ explicit schema or data model, formalized in DDL, XSD, DDI, RDFS, JSON-Schema, data-package or similar
- ◉ community-shared schema or data model , available from a standard location

8. Comprehensible - supported with unambiguous definitions for all internal elements

- ○ local field codes or labels
- ○ labels with full text explanations
- ○ community standard labels (e.g. CF Conventions, UCUM units)
- ○ some fields linked to externally managed definitions
- ◉ all fields linked to standard, externally managed definitions

9. Linked - to other data and definitions using public identifiers (e.g. URIs)

- ○ no links
- ○ in-bound links from a catalogue or landing-page
- ◉ out-bound links to related data and definitions

10. Licensed - conditions for re-use are available and clearly expressed

- ○ no license
- ○ license described in text
- ◉ link to a standard license (e.g. Creative Commons)

Tell us about your data

... maintenance and provenance

11. Curated - commitment to ensuring the data is available long term

- ○ once-off dump, no ongoing commitment
- ◉ best effort, project website
- ○ public or institutional repository (e.g. CKAN, GitHub)
- ○ certified repository

12. Updated - part of a regular data collection program or series, with clear maintenance arrangements and update schedule

- ○ one-time dataset
- ◉ part of series - occasional/irregular update
- ○ part of series - regular scheduled updates

13. Assessable - accompanied by, or linked to, a data-quality assessment and description of the origin and workflow that produced the data

- ◉ no quality or lineage information
- ○ text lineage statement
- ○ formal provenance trace (e.g. PROV-O)

14. Trusted - accompanied by, or linked to, information about how the data has been used, by whom, and how many times

- ◉ no information about usage
- ○ usage statistics available
- ○ Clearly endorsed by reputable organization or framework

Tell us about your data

Project, organisational, institutional

15. * Complexity of the project

- ○ low
- ○ medium
- ◉ high

16. Cross-organisational project?

- ○ 1 organisation
- ○ 2-4 organisations
- ◉ 5 or more organisations

The detailed answers to the questionnaire about the TCS Datasets or Services Metadata are:

Questionaire

Tell us about your data

... publication and indexing

1. * Dataset identity

Dataset name or title	EPOS TCS Median Dataset or Service
URL	http://ics-c.epos-ip.org/

2. * Published - is the data accessible to users other than the creator or owner?

- ○ No
- ○ By individual arrangement
- ○ File download
- ○ Institutional or community repository
- ◉ Bespoke web service (informal API)
- ○ Bespoke web service (OpenAPI/Swagger)
- ○ Standard web service API (e.g. OGC)

3. Citeable - denoted using a formal identifier

- ○ Not citeable
- ○ Local identifier
- ◉ Web address (URL - not guaranteed stable)
- ○ Persistent web identifier (URI)

4. Described - tagged with metadata

- ○ No metadata
- ○ Abstract and keywords
- ○ Basic metadata (e.g. Dublin Core)
- ○ Specialized metadata (e.g. Darwin Core, ISO 19115/19139, schema.org scientific data profile)
- ◉ Rich metadata using multiple standard RDF vocabularies (e.g. DCAT, PROV, ADMS, GeoDCAT, FOAF, ORG, GeoSPARQL)

5. Findable - indexed in a discovery system

- ○ no
- ○ local or internal system only
- ◉ community wide or jurisdictional system
- ○ highly ranked in general purpose index (Google, Bing etc)

Tell us about your data

... linked and useable

6. Loadable - represented using a common or community-endorsed (i.e. standard) format

- ○ bespoke format (text, binary)
- ◉ one standard format, denoted by a MIME-type
- ○ multiple standard formats

7. Useable - structured using a discoverable, community-endorsed (standard?) schema or data model

- ○ no formal schema
- ◉ explicit schema or data model, formalized in DDL, XSD, DDI, RDFS, JSON-Schema, data-package or similar
- ○ community-shared schema or data model , available from a standard location

8. Comprehensible - supported with unambiguous definitions for all internal elements

- ○ local field codes or labels
- ◉ labels with full text explanations
- ○ community standard labels (e.g. CF Conventions, UCUM units)
- ○ some fields linked to externally managed definitions
- ○ all fields linked to standard, externally managed definitions

9. Linked - to other data and definitions using public identifiers (e.g. URIs)

- ○ no links
- ◉ in-bound links from a catalogue or landing-page
- ○ out-bound links to related data and definitions

10. Licensed - conditions for re-use are available and clearly expressed

- ○ no license
- ◉ license described in text
- ○ link to a standard license (e.g. Creative Commons)

Tell us about your data

... maintenance and provenance

11. Curated - commitment to ensuring the data is available long term

- ○ once-off dump, no ongoing commitment
- ◉ best effort, project website
- ○ public or institutional repository (e.g. CKAN, GitHub)
- ○ certified repository

12. Updated - part of a regular data collection program or series, with clear maintenance arrangements and update schedule

- ○ one-time dataset
- ◉ part of series - occasional/irregular update
- ○ part of series - regular scheduled updates

13. Assessable - accompanied by, or linked to, a data-quality assessment and description of the origin and workflow that produced the data

- ○ no quality or lineage information
- ◉ text lineage statement
- ○ formal provenance trace (e.g. PROV-O)

14. Trusted - accompanied by, or linked to, information about how the data has been used, by whom, and how many times

- ◉ no information about usage
- ○ usage statistics available
- ○ Clearly endorsed by reputable organization or framework

Tell us about your data

Project, organisational, institutional

15. * Complexity of the project

- ○ low
- ○ medium
- ◉ high

16. Cross-organisational project?

- ○ 1 organisation
- ○ 2-4 organisations
- ◉ 5 or more organisations

Annex C – yaml files

```
survey:
  date: 2019-04-03
  version: 1
  creator:
    name: Luca Trani
    email: trani@knmi.nl
infrastructure:
  acronym: EPOS
  name: EPOS ORFEUS
  recognized authority URL: NULL
  domain: earth
  repositories:
  - URL: http://orfeus-eu.org/webdc3/
    name: European Federated Data Archive
    kind: metadata data repository
    data repository type: domain
    metadata repository type: various
    software: GitHub
    identifier:
    - IRI: https://www.doi.org/
      kind: DOI
      system: DataCite
      assigned: manually
      provider: DataCite
      includes-attributes:
      - file location
      - file checksum
      - PID replica
    - IRI: https://www.pidconsortium.eu/
      kind: PID
      system: ePIC
      assigned: automatically
      provider: ePIC service
      includes-attributes: none
    certification methods: NULL
    policies:
    - usage
    - acknowledgements
    - citation
    registries:
    - local registry
    persistency-guaranty: none
    access mechanisms:
      authentication method: B2ACCESS
      access protocol URL: NULL
      access without costs: NULL
      own user database maintained: yes
      ORCID used in AAI: yes
      major access technology supported: HTTP
      authorisation technique: NULL
      authorisation needed for: NULL
      authorization for accessing content needed: NULL
      data licenses in use:
      - CC BY
      data license IRI: NULL
      metadata openly available: yes
    data:
    - type name: seismic waveform
      preferred formats:
      - format name: MiniSEED
        metadata types in data headers: none
      registered data schema: yes
      search on data: yes
      search engine indexing: NULL
    - type name: quake event
      preferred formats:
      - format name: quakeML
        metadata types in data headers: none
      registered data schema: no
```

```
    search on data: yes
    search engine indexing: NULL
  metadata:
    schema:
    - URL: NULL
      name: stationXML
      provenance fields included: none
    categories defined in registries: yes
    PIDs included: partially
    primary storage format: XML
    metadata longevity plan URL: NULL
    format IRI: NULL
    export formats supported:
    - XML
    - SEED
    exchange/harvesting methods: NULL
    local search engine URL: NULL
    external search engine types supported: NULL
    access policy statements included: NULL
    machine actionable: NULL
  vocabularies:
    IRI: NULL
    type: NULL
    topic: NULL
    name: NULL
    specification language URL: NULL
    concept IRI: NULL
  data management plans:
    specific DMP tools used: DMPonline
    data publishing steps applied: NULL
  data processing:
    special data processing steps applied: metadata extraction
    workflow frameworks applied:
    - VERCE
    - DARE
    distributed workflows tools used:
    - VERCE
    - DARE
    other analysis services offered: none
    data products offered: NULL
  fairness:
    data findability:
      data findable: yes
      gaps: none
    data accessibility:
      data accessible: yes
      gaps: none
    data interoperability:
      data interoperable: partially
      gaps: improving vocabularies
    data re-usability:
      data reusable: yes
      gaps: improving vocabularies
  test fairness:
    URL/IRI of dataset: http://www.orfeus-eu.org/data/eida/
    URL of discovery portal: http://orfeus-eu.org/webdc3/
    IRI of machine readable metadata of dataset: NULL
    machine readable provenance: NULL
    compliance validation: NULL
```

Annex D – SPARQL example

Below, results of a query on answers related to I1 FAIR principle are shown. From this overview it is clear that almost all the contributing RIs make use of standard metadata languages.
This example comes from the Marine Domain, since the process for the Solid Earth domain is still ongoing.

Overviews (Example: answers around I1..) (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation

	l	p	o
0	SDN Data F	http://envri.eu/ns/categoriesAreDefinedInRegistries	1.23 partially
1	LifeWatch ,	http://envri.eu/ns/categoriesAreDefinedInRegistries	true
2	ICOS Carbo	http://envri.eu/ns/categoriesAreDefinedInRegistries	true
3	SDN SDN C	http://envri.eu/ns/categoriesAreDefinedInRegistries	true
4	LifeWatch I	http://envri.eu/ns/categoriesAreDefinedInRegistries	true
5	LifeWatch I	http://envri.eu/ns/categoriesAreDefinedInRegistries	true
6	Euro-Argo I	http://envri.eu/ns/categoriesAreDefinedInRegistries	planned
7	Euro-Argo I	http://envri.eu/ns/hasVocabularyIri	http://seadatanet.maris2.nl/v_bodc_vocab_v2/search.asp?lib=P06
8	SDN SDN C	http://envri.eu/ns/hasVocabularyIri	http://standards.iso.org/ittf/PubliclyAvailableStandards/ISO_19139_Schemas/resou
9	ICOS Carbo	http://envri.eu/ns/hasVocabularyIri	http://purl.org/dc/elements/1.1/
10	LifeWatch ,	http://envri.eu/ns/hasVocabularyIri	http://www.marinespecies.org/
11	SDN Data F	http://envri.eu/ns/hasVocabularyIri	http://vocab.nerc.ac.uk/collection/V22/current/
12	SDN Data F	http://envri.eu/ns/hasVocabularyIri	http://standards.iso.org/ittf/PubliclyAvailableStandards/ISO_19139_Schemas/resou
13	LifeWatch I	http://envri.eu/ns/hasVocabularyIri	http://www.marineregions.org/
14	ICOS Carbo	http://envri.eu/ns/hasVocabularyIri	http://www.w3.org/ns/prov
15	SDN SDN C	http://envri.eu/ns/hasVocabularyIri	https://www.bodc.ac.uk/resources/vocabularies/vocabulary_search/
16	Euro-Argo I	http://envri.eu/ns/hasVocabularyIri	http://seadatanet.maris2.nl/v_bodc_vocab_v2/search.asp?lib=P06
17	ICOS Carbo	http://envri.eu/ns/hasVocabularyIri	http://meta.icos-cp.eu/ontologies/cpmeta/
18	LifeWatch I	http://envri.eu/ns/hasVocabularyIri	https://www.bodc.ac.uk/resources/vocabularies/vocabulary_search/
19	LifeWatch I	http://envri.eu/ns/hasVocabularyIri	http://www.marinespecies.org/
20	LifeWatch I	http://envri.eu/ns/hasVocabularyIri	http://www.marineregions.org/
21	LifeWatch I	http://envri.eu/ns/hasVocabularyIri	http://www.marinespecies.org/

Appendix A - Glossary

Acronyms and abbreviations

CERIF	Common European Research Information Format
ENVRI	Environmental Research Infrastructure
EOSC	European Open Science Cloud
ESFRI	European Strategy Forum on Research Infrastructures
IRI	Internationalized Resource Identifiers
OGC	Open Geospatial Consortium
PROV	W3C PROV family of documents (https://www.w3.org/TR/prov-overview/)
RDF	Resource Description Framework
URI	Uniform Resource Identifier
W3C	World Wide Web Consortium