EPOS: A FAIR Research Infrastructure

Keith G Jeffery
Keith G Jeffery Consultants
Faringdon, UK
Email: keith.jeffery@keithgjefferyconsultants.co.uk

Daniele Bailo
ERIC
Istituto Nazionale di Geofisica e Vulcanologia
Rome, Italy
Email: daniele.bailo@ingv.it

Kuvvet Atakan
Department of Earth Science
University of Bergen
Bergen, Norway
Email: kuvvet.atakan@uib.no

Matt Harrison
Director Informatics
British Geological Survey
Keyworth, UK
Email: mharr@bgs.ac.uk

***Abstract*—The European Plate Observing System (EPOS) has been developed over some years and is now in transition to full operational status. It currently offers a portal with access to more than 200 data services, the portals of constituent research communities and prototype access to a service for Trans-National Access (TNA) to equipment and sensors as well as access to information on the organizations and persons involved in EPOS together with research capabilities. From the beginning, EPOS was designed to support the FAIR (Findable, Accessible, Interoperable, Reusable) principles. This paper explains how the EPOS architecture meets the specifications of FAIRness.**

Keywords- geoscience; data services; metadata; CERIF; catalog; research infrastructures; FAIR.

I. INTRODUCTION

The architecture of EPOS was described in [1]. The present work focuses on how FAIR principles are applied in EPOS. The purpose of EPOS is to provide end-users – including researchers, educators, policymakers, industry employees, citizen scientists – with the ability to discover, contextualize and utilize the heterogeneous assets of the various geoscience communities through a homogeneous interface.

A. Overview

The architecture has been designed to satisfy the following criteria:

1. Minimal interference with existing communities' operations and developments, including Information Technology (IT);
2. Easy-to-use user interface;
3. Access to assets through a metadata catalog: initially services, but progressively also datasets, workflows, software modules, computational facilities, instruments/sensors, all with associated organizational information including experts and service managers;
4. Progressive assistance in composing workflows of services, software and data to deploy on e-Infrastructures to achieve research infrastructure user objectives.

B. FAIRness

From the beginning, EPOS was designed to be FAIR and EPOS participants were involved in the discussions leading to the FAIR principles [2] and also subsequent work on FAIR metrics within the FAIR Data Maturity Model Working Group of Research Data Alliance (RDA) [3]. The major contributions of this paper are to indicate (a) how FAIRness was achieved from the beginning of EPOS: (b) how our systems development approach maps to the FAIR principles using a 'pyramid' diagram.

C. Previous Work

EPOS provides an original approach to the provision of homogeneous access over heterogeneous digital assets and providing FAIRness. Previous work on homogenizing heterogeneity has been mainly within a limited domain (where standards for assets and their metadata may be consensual across the whole domain thus reducing heterogeneity) with manual processes and associated costs. Filematch [4] exhibited those problems. NASA has a Common Metadata Repository (CMM). In 2013, NASA developed the Unified Metadata Model (UMM) [5] to and from which, other metadata standards are converted. This follows the superset canonical rich metadata approach already used in EPOS. The Open Geospatial Consortium (OGC) has produced a series of standards. GeoNetwork [6] has established a suite of software based around the OGC ISO19115 metadata standard; however, despite its open nature, this software 'locks in' the developer to a particular way of processing, does not assist in the composition and deployment of workflows and the metadata is insufficiently rich for automated processing. EarthCube [7] is a collection of projects providing designs and tools for

 ISBN: 978-1-61208-762-7

geoscience, including interoperability, in USA. The project encountered – by using pairwise brokering – the problem that it required n*(n-1) brokers instead of the n required if a canonical metadata approach is used. Auscope [8] includes AuScope GRID which, by using ISO19115, encounters the problems outlined above. GEOSS [9] uses the 'system of systems' approach, but this requires many bilateral interfaces with the combinatorial problem discussed above.

In essence, all these other approaches provide some degree of FAIRness (Finding, Accessing), but usually require human and manual work to achieve interoperability or reuse.

EPOS, with its superset rich canonical metadata, overcomes the problems concerning homogeneous access over heterogeneous assets and, furthermore, provides increasingly automated FAIRness.

The rest of the paper is organized as follows: Section II describes the architecture; Section III discusses the importance of metadata; Section IV demonstrates that EPOS is FAIR and Section V summarizes conclusions.

II. ARCHITECTURE

The Information and Communication Technologies (ICT) architecture of EPOS is designed to facilitate the research community and others in discovering and utilizing through the Integrated Core Services (ICS) the assets provided by the Thematic Core Services (TCS) communities. The architecture was described in [1], but is recapitulated briefly for this paper.

A. Introduction

In order to provide end-users with homogeneous access to services and multidisciplinary data collected by monitoring infrastructures and experimental facilities (and to software, processing and visualization tools as well), a complex, scalable and reliable architecture is required. A diagram of the architecture is outlined in Figure 1.

The key aspects are:

1. National Research Infrastructures (NRI) hold the assets and provide metadata to describe them;
2. Thematic Core Services (TCS) that relate to (currently 10) communities, each for a particular domain of geoscience. These communities harmonise progressively semantic aspects of metadata such as terminology in ontologies and also decide which NRI assets should be proposed for availability through EPOS;
3. Integrated Core Services (ICS) that provide the portal, associated metadata catalog and thus provide Findability, Accessibility, Interoperability, and Reusability (FAIR).

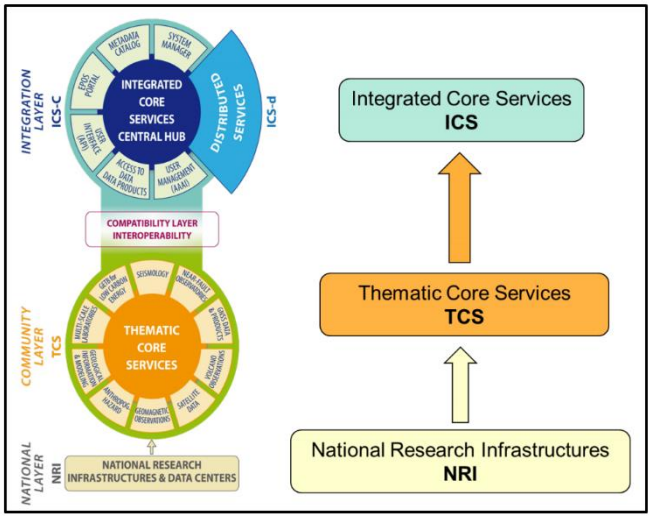

Figure 1. EPOS Architecture.

B. ICS

The EPOS-ICS provides the entry point to the EPOS environment. ICS-C provides the portal and metadata catalog, with associated convertors, to accept metadata from TCS and ingest into the catalog. ICS-D provides distributed computational resources including also processing and visualization services, of which a specialization is Computational Earth Science (CES). ICS-C provides the basis for deployment of workflows, including to ICS-D facilities, that in turn rely on e-Infrastructures such as Cloud Computing or supercomputing. EPOS has also been involved in the VRE4EIC project [10] (and cooperating with EVER-EST [11]) to ensure convergent evolution of the EPOS ICS-C user interface and Application Programming Interfaces (APIs) for programmatic access with the developing Virtual Research Environments (VREs). EPOS participates in the recently approved ENVRIFAIR project [12] that will improve the deployments to the European Open Science Cloud (EOSC) [13] (See Figure 2).

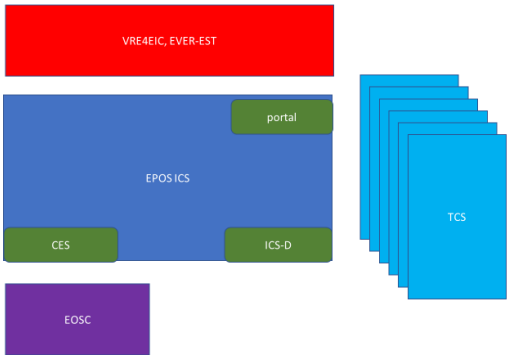

Figure 2. EPOS Positioning.

 ISBN: 978-1-61208-762-7

Workflow for the deployment (which may be a simple file download or a complex set of services including analytics and visualization) will be generated within the ICS-C, by interaction with the users. The workflow will be checked by the end-user before deployment. However, the detailed content/capability of the assets might not be known, e.g., the dataset may not contain the relevant information despite its metadata description, or the software may not execute as the user expects despite the metadata description. The execution of the deployment is monitored and execution information is returned to the end-user. The ICS represents to the end-user the infrastructure, consisting of services that will allow access to multidisciplinary resources provided by the TCS. These will include data and data products as well as synthetic data from simulations, processing, and visualization tools. The key to this view of the geoscience domain is the metadata catalog using the Common European Research Information Format (CERIF) [14].

C. ICS-C

The ICS-C consists of multiple logical areas of functionality, these include the Graphical User Interface (GUI), web-API, metadata catalogue, user management etc. A micro-service architecture has been adopted in the ICS-C, where each (micro) services is atomic and dedicated to a specific class of tasks. The EPOS ICS-C system architecture is outlined in (Figure 3). The Microservices architecture approach envisages small atomic services dedicated to the execution of a specific class of tasks, which have high reliability [15][16]. Docker Containers technology was used. enabling complete isolation of independent software applications running in a shared environment. The communication between microservices is done via messages received and sent on a queueing system, in this case RabbitMQ [17]. As a result, a chain of microservices processes the requests.

The current architecture includes an Authentication, Authorization, Accounting Infrastructure (AAAI). This has been implemented using UNITY [18] and has involved close cooperation with CYFRONET, evolving to the integrated authentication system for research communities. Authorization is more complex, and is being developed incrementally, as it depends on rules agreed with the TCS (within the context of the financial, legal and governance traversal workpackages of EPOS-IP) for each of their assets, and included further metadata elements into the CERIF catalog to control such authorization. The latter has been prepared and awaits validation by the TCS. Related to this, the GUI now provides a user notification pointing to a legal disclaimer for the EPOS system. It should be noted that use of authentication and authorization does not preclude FAIRness, but does allow for protection of assets e.g. to allow a research team time to publish results based on their data before the data is made generally available.

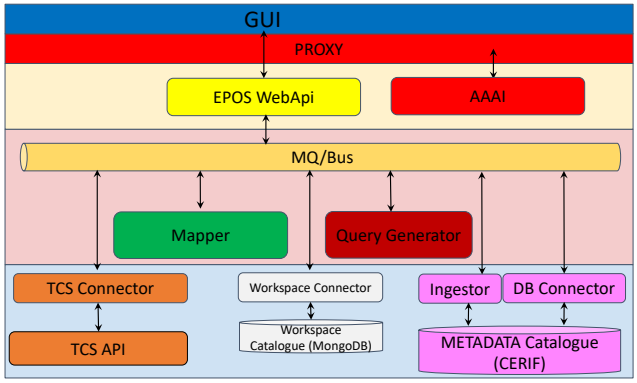

Figure 3. ICS-C Architecture.

The topic of workflow has required pilot projects with TCS experts to clarify the requirements, available technologies and the difficulties of appropriate user interactions. Working in cooperation with the VRE4EIC project we have the basic components for (a) a general workflow manager interface; (b) interfaces to specific workflow managers such as Taverna [19].

D. ICS-D

ICS-D concerns workflow management, since once Found and Accessed, the assets are Interoperable and Reusable, using workflows distributed across e-Infrastructure components by ICS-D. A specification of the metadata elements required for ICS-D has been developed, and is still being refined in the light of experience from the pilots mentioned above. ICS-D will appear to the workflow, or to the end-user, as a service accessed through an API. The deployment requires middleware. Results from the PaaSage project [20] are relevant and the concurrent MELODIC project [21] offers optimization, including that based on dataset placement and latency. Further refinement of requirements and the architectural interfaces continues.

III. METADATA

The core of the EPOS architecture is the metadata catalog and specifically the superset, rich, canonical metadata format chosen, namely CERIF. This allows EPOS to provide support for cross-domain, interoperable science while achieving the objectives of the FAIR principles.

A. Introduction

The metadata catalogue is the way of representing in a homogeneous way, the heterogeneous assets provided within the EPOS community. The catalog defines what assets are visible to end-users. It provides the required information to facilitate Finding, Accessing, Interoperating and Reusing (FAIR) EPOS assets. In fact, between Finding and Accessing, the use of a rich format like CERIF also allows contextualization: that is the assessment of relevance and quality of the asset for the purpose in hand. Furthermore, the use of linking entities between base entities in CERIF, with

 ISBN: 978-1-61208-762-7

role and temporal interval, provides automatically records describing provenance since it is possible to retrieve all link entity records related to a particular base entity, role or time interval in any combination. The catalogue contains: (i) technical specification to enable autonomic ICS access to TCS discovery and access services, (ii) metadata associated with the digital object with a direct link to it, (iii) information about users, resources, software, and services other than data services (e.g., rock mechanics, geochemical analysis, visualization, processing).

The CERIF data model was chosen because it: (1) separates base entities from linking entities, thus providing a fully connected graph structure; (2) using the same syntax, stores the semantics associated with values of attributes, both for base entities and (for role of the relationship) for linking entities, that also store the temporal duration of the validity of the linkage. This provides great power and flexibility. CERIF also (as a superset) can interoperate with widely adopted metadata formats such as Dublin Core (DC) [22], Data Catalogue Vocabulary (DCAT) [23], Comprehensive Knowledge Archive Framework (CKAN) [24], INSPIRE (the EC version of ISO 19115 for geospatial data) [25] and others using convertors developed as required to meet the metadata mappings achieved between each of the above standards and CERIF. Currently 17 different metadata formats in geoscience are convertible with CERIF. The metadata catalogue also manages the semantics, in order to provide the meaning of the attribute values.

To recap, the use of CERIF automatically provides:

(a) The ability for discovery, contextualization, interoperation and (re-)use of assets according to the FAIR principles [2]
(b) A clear separation of base entities (things) from link entities (relationships);
(c) Formal syntax and declared semantics;
(d) A semantic layer, also with the base/link structure allowing crosswalks between semantic terminology spaces;
(e) Conversion to/from other common metadata formats;
(f) Built-in provenance information, because of the timestamped role-based links;
(g) Curation facilities, because of being able to manage versions, replicates and partitions of digital objects using the base/link structure;

These technical properties of CERIF provide that which is required to ensure FAIRness of the system. The catalog is constantly evolving with the addition of new assets (such as services, datasets), but also increasingly rich metadata, as the TCSs improve their metadata collection to enable more autonomic processing.

B. TCS Metadata

The 'treasure' of EPOS is the assets provided, through the TCS communities from the NRIs. These TCS Data, Data Products, Software and Services (DDSS) are described by metadata. The metadata describing those assets is supplied via the TCS IT experts and is harmonized as much as possible. It is checked for quality, and registered in the granularity database (see below). This relates to governance, including funding for the TCS. It is then converted to CERIF via an intermediate format (see below).

C. ICS Metadata

The intermediate format is known as the EPOS baseline. It provides a minimum set of common metadata elements required to operate the ICS, taking into consideration the heterogeneity of the assets of the many TCSs involved in EPOS. It has been implemented as an application profile using an extension of the DCAT standard, namely the EPOS-DCAT-AP. The baseline can be extended to accommodate extra metadata elements, where it is deemed that those metadata elements are critical in describing and delivering the data services for any given community. Indeed, this has happened already when the original EPOS-DCAT-AP was found to be inadequate, and a new version with richer metadata was designed and implemented.

The metadata to be obtained from the EPOS TCSs, as described in the baseline document (and any other agreed elements) will be mapped to the EPOS ICS CERIF catalog. The process of converting metadata acquired from the EPOS TCS to CERIF will be done by in consultation with each TCS as to what metadata they have available and harvesting mechanisms

The metadata is ingested from the TCS community NGIs by various mechanisms, depending on local conditions. In general, they expose an API allowing the metadata to be collected. The metadata is transformed from local format to EPOS baseline and thence to CERIF. These APIs, and the corresponding ICS convertors, collectively form the "interoperability layer" in EPOS, which is the link between the TCSs and the ICS.

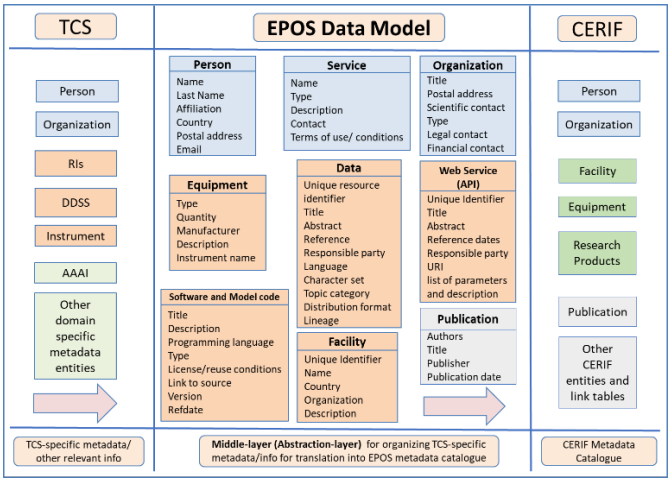

Figure 4. EPOS Metadata Baseline.

The EPOS baseline can thus be considered as an intermediate layer, that facilitates the conversion from the

community metadata standards such as ISO19115/19, DCAT, Dublin Core, INSPIRE, etc., describing the DDSS elements and not the index or detailed scientific data (See Figure 4).

D. DDSS and Granularity Database

As a part of the Requirements and Use cases Collection (RUC) from the TCSs, a specific list was prepared to include all data, data products, software and services (DDSS). The DDSS master table was originally implemented as Excel spreadsheets. The DDSS Master Table was also used for extracting the level of maturity of the various DDSS elements in each TCS, as well as providing a summary of the status of the TCS preparations for the ICS integration and interoperability. The current version of the DDSS Master Table consists of 363 DDSS elements, where 165 of these already exist and are declared by TCSs to be ready for implementation. The remaining DDSS elements required more time to harmonize the internal standards, prepare an adequate metadata structure and so are available for implementation soon. In total, 21 different harmonization groups (HGs) are established to help organizing the harmonization issues in a structured way. In addition, user feedback groups (UFGs) have been established and work to give constant and structured feedback during the implementation process of the TCS-ICS integration and the development of the ICS.

The rate of change of the DDSS maser table indicated that a different technology should be used. The DDSS master table has been transformed to the granularity database because of the problems of referential and functional integrity using a spreadsheet; relational technology provides appropriate constraints to ensure integrity.

An increasingly detailed RUC collection process is formulated and explained through dedicated guidelines and interview templates. A roadmap for the ICS-TCS interactions for the RUC collection process was prepared for this purpose and distributed to all TCSs.

In this approach, a five-step procedure is applied involving the following:

- Step 1: First round of RUC collection for mapping the TCS assets;
- Step 2: Second round of RUC collection for identifying TCS priorities;
- Step 3: ICS-TCS Integration Workshop for building a common understanding for metadata
- Step 4: Third round of RUC collection for refined descriptions before implementation;
- Step 5: Implementation of RUC to the CERIF metadata;

This procedure has been refined over man months, but is designed to ensure maximum richness, integrity and correctness of the metadata, since it is upon the quality of the metadata that the achievement of FAIRness depends.

Work is now complete in converting the DDSS tables (in Excel) to the granularity database using Postgres. This (a) facilitates finding particular DDSS elements, eliminating duplicates and checking the progress of getting DDSS elements into the metadata format; (b) simplifies harvesting to the metadata catalog.

IV. DEMONSTRATING THAT EPOS IS FAIR

The mapping of the FAIR principles to aspects of the EPOS architecture, demonstrates that the FAIR principles are supported by the EPOS architecture from metadata to service provision (Figure 5).

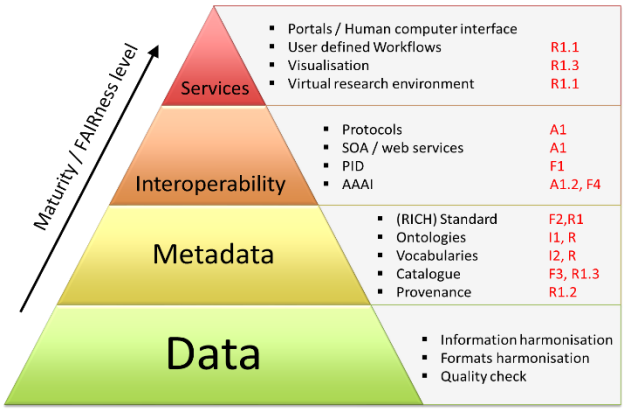

Figure 5. The FAIR Principles and the EPOS Architecture Pyramid [27] [28].

The provision of FAIRness starts with the metadata as explained above. To achieve FAIRness, the metadata must be rich (many attributes), identify uniquely the asset with a Resolvable Universally Unique Persistent IDentifier (RUUPID), have available licensing information, use standard protocols, have an appropriate vocabulary, provide qualified references and provenance. We believe to this should be added demonstrate both referential and functional integrity. It is on the latter two quality measures that many other metadata formats fail.

Findability is achieved by the rich metadata. Query on the rich metadata selects the metadata representing the assets of interest, including the RUUPID of the asset.

Accessibility is achieved by resolving the asset RUUPID and also ensuring the access conditions – in a licence (better a machine-representation of the conditions in the licence) or metadata concerning authorization from the AAAI – are respected.

Interoperability is achieved by the use of convertors between metadata formats, to provide homogeneous access to the assets through standard APIs. If necessary, data formats can also be converted to a canonical form to allow co-analysis or display of heterogeneous datasets.

Reusability is achieved because of the richness of the metadata (many attributes), the provision of licence information allied to the authorization component of AAAI, the utilization of community standard formats and finally the provision of provenance information which comes automatically because of the time-stamped role-based relationships between base entities in CERIF.

 ISBN: 978-1-61208-762-7

V. CONCLUSION

Currently, 186 digital assets (rising progressively to 221) from the domain communities, supported by 281 webservices, are represented by CERIF metadata in the EPOS ICS-C catalog and made available FAIRly. These services, described by the metadata, can be discovered, accessed, contextualized and (re)utilized individually or composed into workflows and hence become interoperable. A GUI provides the user view onto the catalog, and it also provides a workspace to collect the metadata of the assets selected for use (Figure 6). From the workspace a workflow may be constructed and deployed.

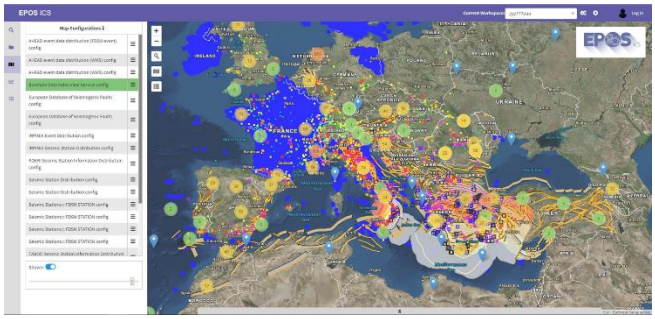

Figure 6. EPOS-ICS Graphical User Interface.

Future plans include:

(a) Harvesting of metadata describing more assets: not only services, but also datasets, software, workflows, equipment;
(b) Improving the GUI to allow workflow deployment with 'fire and forget' technology, or single-step with user checking and adjustment at each step;
(c) Completion of the software to permit trans-national access to laboratory and sensor equipment;
(d) Improved AAAI to give the domain communities finer control over FAIR utilisation of their assets;
(e) The inclusion of virtual laboratory-type interfaces (virtual research environments), allowing users access and connectivity including open-source frameworks such as Jupyter notebooks [26], which are increasingly being used in some scientific communities.

The architecture outlined and demonstrated (in successive prototypes) in EPOS-IP has found favour (not without some criticism of course – leading to agile improvements) from the user community. The criticisms usually concerned: (a) simplifying the complexity of the user interface (achieved by the use of panes); (b) improvements in the quantity (more attributes) and quality of metadata to make Finding, Accessing Interoperating and Reusing easier – this was really a criticism of the TCS supplied metadata more than the ICS; (c) lack of harmonization – again this is the responsibility of harmonization groups across the TCS communities. Furthermore, the prototype system has passed Technological Readiness Assessment procedures within the governance of the EPOS-IP project. Currently the ICS is undergoing pre-production tests. The architecture meets the requirements, it is state of the art and has a further development plan. The FAIR achievements are:

1. EPOS architecture from the beginning was designed for FAIR, with EPOS staff involved in FAIR definition and subsequent indicators work;
2. EPOS is already FAIR-compliant with RUUPIDs, rich metadata (many attributes), formal syntax, declared semantics, referential and functional integrity;
3. The EPOS catalog already interoperates with 17 metadata 'standards' in geoscience and wider;
4. EPOS is open to interoperate with other RIs
(a) directly; (b) via an 'umbrella' VRE; or (c) via EOSC;
5.EPOS started with interoperable services which overcomes many problems with data and is anticipating EOSC.

ACKNOWLEDGMENT

The authors acknowledge the work of the whole ICT team in EPOS reported here, and the funding of the European Commission H2020 program (Grant agreement 676564) and National Funding Councils that have made this work possible.

REFERENCES

[1] K. Jeffery, D. Bailo, K. Atakan, and M. Harrison "EPOS: European Plate Observing System" in Proc. Eleventh International Conference on Advanced Geographic Information Systems, Applications, and Services (GEOProcessing 2019), pp. 79-86.
[2] FAIR Principles https://www.force11.org/group/fairgroup/fairprinciples (accessed on 30 January 2020)
[3] RDA Working Group https://www.rd-alliance.org/groups/fair-data-maturity-model-wg (accessed on 30 January 2020)
[4] P. Sutterlin, K. Jeffery, and E. Gill: "Filematch: A Format for the Interchange of Computer-Based Files of Structured Data" Computers and Geosciences 3 1977) pp. 429-468.
[5] UMM: https://earthdata.nasa.gov/about/science-system-description/eosdis-components/common-metadata-repository/unified-metadata-model-umm (accessed on 30 January 2020)
[6] Geonetwork https://geonetwork-opensource.org/ (accessed 30 May 2019)
[7] EarthCube: https://www.earthcube.org/ (accessed on 30 January 2020)
[8] AuScope: http://www.auscope.org.au/ (accessed on 30 January 2020)
[9] GEOSS: https://www.earthobservations.org/geoss.php (accessed on 30 January 2020)9)
[10] VRE4EIC: https://www.vre4eic.eu/ (accessed on 30 January 2020)
[11] EVEREST: https://ever-est.eu/ (accessed on 30 January 2020)
[12] ENVRI-FAIR: http://envri.eu/envri-fair/ (accessed on 30 January 2020)
[13] EOSC: https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud (accessed on 30 January 2020)

 ISBN: 978-1-61208-762-7

[14] CERIF: https://www.eurocris.org/cerif/main-features-cerif (accessed on 30 January 2020)
[15] Newman, Sam. “Building Microservices”, O'Reilly Media, Inc., 2015
[16] International Journal of Open Information Technologies ISSN: 2307- 8162
[17] RabbitMQ: https://www.rabbitmq.com/ (accessed on 30 January 2020)
[18] UNITY: http://www.unity-idm.eu (accessed on 30 January 2020)
[19] Taverna: https://taverna.incubator.apache.org/ (accessed on 30 January 2020) [20] PaaSage: https://paasage.ercim.eu/ (accessed on 30 January 2020)
[21] MELODIC: melodic.cloud/ (accessed on 30 January 2020)
[22] DC: http://dublincore.org/documents/dces/ (accessed on 30 January 2020)
[23] DCAT: https://www.w3.org/TR/vocab-dcat/ (accessed on 30 January 2020)
[24] CKAN: https://ckan.org/ (accessed on 30 January 2020)
[25] INSPIRE: https://inspire.ec.europa.eu/ (accessed on 30 January 2020)
[26] Jupyter: https://jupyter.org/ (accessed on 30 January 2020)
[27] D. Bailo, (2019, July 10). “Four-stages FAIR Roadmap - FAIR "Pyramid"”. Zenodo. http://doi.org/10.5281/zenodo.3299353 (accessed on 30 January 2020)
[28] D. Bailo, R. Paciello, M. Sbarra, R. Rabissoni, V. Vinciarelli and M. Cocco “Perspectives on the Implementation of FAIR Principles in Solid Earth Research Infrastructures” , Frontiers in Earth Science, vol 8, 2020, p.3, DOI10.3389/feart.2020.00003 https://www.frontiersin.org/article/10.3389/feart.2020.00003 (accessed on 30 January 2020)

 ISBN: 978-1-61208-762-7