ANNALS OF GEOPHYSICS, 65, 2, DM210, 2022; doi:10444.01/ag-8742
OPEN ACCESS

Data integration and FAIR data management in Solid Earth Science

Daniele Bailo[*,1], Keith G. Jeffery[2], Kuvvet Atakan[3], Luca Trani[4], Rossana Paciello[1], Valerio Vinciarelli[5], Jan Michalek[3], Alessandro Spinuso[4]

(1) Istituto Nazionale di Geofisica e Vulcanologia (INGV), Rome, Italy
(2) Keith G Jeffery Consultants, United Kingdom
(3) Universitetet i Bergen (UiB) 5020 Bergen, Norway
(4) Department of R&D Seismology and Acoustics, Royal Netherlands Meteorological Institute (KNMI), Utrechtseweg 297, 3731 GA, De Bilt, The Netherlands.
(5) EPOS ERIC, Via di Vigna Murata 605, Rome, Italy

Article history: received October 28, 2021; accepted March 9, 2022

Abstract

Integrated use of multidisciplinary data is nowadays a recognized trend in scientific research, in particular in the domain of solid Earth science where the understanding of a physical process is improved and made complete by different types of measurements – for instance, ground acceleration, SAR imaging, crustal deformation – describing a physical phenomenon. FAIR principles are recognized as a means to foster data integration by providing a common set of criteria for building data stewardship systems for Open Science.

However, the implementation of FAIR principles raises issues along dimensions like governance and legal beyond, of course, the technical one. In the latter, in particular, the development of FAIR data provision systems is often delegated to Research Infrastructures or data providers, with support in terms of metrics and best practices offered by cluster projects or dedicated initiatives.

In the current work, we describe the approach to FAIR data management in the European Plate Observing System (EPOS), a distributed research infrastructure in the solid Earth science domain that includes more than 250 individual research infrastructures across 25 countries in Europe. We focus in particular on the technical aspects, but including also governance, policies and organizational elements, by describing the architecture of the EPOS delivery framework both from the organizational and technical point of view and by outlining the key principles used in the technical design. We describe how a combination of approaches, namely rich metadata and service-based systems design, are required to achieve data integration. We show the system architecture and the basic features of the EPOS data portal, that integrates data from more than 220 services in a FAIR way. The construction of such a portal was driven by the EPOS FAIR data management approach, that by defining a clear roadmap for compliance with the FAIR principles, produced a number of best practices and technical approaches for complying with the FAIR principles.

Such a work, that spans over a decade but concentrates the key efforts in the last 5 years with the EPOS Implementation Phase project and the establishment of EPOS-ERIC, was carried out in synergy with other EU initiatives dealing with FAIR data.

On the basis of the EPOS experience, future directions are outlined, emphasizing the need to provide i) FAIR reference architectures that can ease data practitioners and engineers from the domain communities to adopt FAIR principles and build FAIR data systems; ii) a FAIR data management

framework addressing FAIR through the entire data lifecycle, including reproducibility and provenance; and iii) the extension of the FAIR principles to policies and governance dimensions.

Keywords: FAIR; Research Infrastructure; Data management; EPOS, multidisciplinary data portal

1. Introduction

In the last decade, integration of distributed resources scattered across regional, national or international data centres and Research Infrastructures (RI), has become a requirement for carrying out multidisciplinary research and fostering innovation in science. This is true in particular for disciplines where various types of measurements are necessary to describe specific physical phenomena, as in the case of solid Earth Science. For instance, the combination of historical seismic sequences, focal mechanisms solutions, InSAR imaging, earthquake localization techniques and differential interferometry from synthetic aperture radar and GPS data, can be used to describe in depth the complexity of fault systems [Chiaraluce, 2009].

Homogenous access to such heterogeneous data sources requires on one hand transparent access to knowledge and data shared by openly accessible providers – which is one of the definitions of "Open Science" [Vicente-Saez et al., 2018] – on the other hand, it requires a considerable technical and engineering undertaking to steward data through open data systems. In this scenario, FAIR guiding principles [Wilkinson et al., 2016] can play a relevant role for data infrastructures providers willing to enable a more effective and open research environment.

FAIR principles are gaining ground and are recognized by the European Commission as fundamental building blocks of the European Open Science Cloud (EOSC)[1] and, more generally, as a mandatory requirement to steward data in an Open Science framework and to participate in EU calls whenever these include management or production of data. However, despite the attempts to support data providers in the hard task of building FAIR data stewardship systems by means of guidelines [Collins et al., 2018; Hong et al., 2020] and supporting initiatives like GO-FAIR [Schultes, 2018], FAIR principles implementation remains largely a business delegated to Research Infrastructures engineers, system architects and data practitioners.

In this landscape, EPOS[2], European Plate Observing System, is a pan-European, large scale research infrastructure within the ESFRI (European Strategic Forum on Research Infrastructures) roadmap[3] that was recently granted the ERIC status, as described in another paper in this special issue.

As a distributed research infrastructure, EPOS fosters the integrated use of data, data products, software, (web) services and facilities from the solid Earth science community in Europe (Figure 1). It includes the national research infrastructures (NRIs) at the bottom of the chain of data provision, through a middle layer – the Thematic Core Services (TCS) – where thematic communities further develop data and products services that are specific for various subdisciplines of solid Earth science, and finally to the integrated core services (ICS), where integrated and interoperable data from 10 thematic communities are provided to various users. NRIs provide data from more than 250 individual research infrastructures installed in 25 countries in Europe, monitoring the European tectonic plate through geographically distributed sensor networks and remote sensing covering the entire European region, as well as other data repositories. TCS are coordinated and governed by 10 TCS consortia which are linked to the EPOS-ERIC through legally binding dedicated collaboration agreements. This framework of FAIR data provision through the TCS and ICS is referred to as the "EPOS delivery framework".

In the current work, we first describe the general EPOS landscape presenting the Thematic Communities and the EPOS architecture. Then, the main concepts underpinning integration of assets as implemented in EPOS are outlined. Subsequently, the EPOS data portal is described together with its main functionalities, demonstrating how the application of FAIR principles supports the integration of multi-disciplinary data and, with the Distributed

1 https://eosc-portal.eu/ (accessed on the 13th of January 2022)

2 https://www.epos-eu.org/ (accessed on the 4th of October 2021)

3 http://roadmap2018.esfri.eu/projects-and-landmarks/browse-the-catalogue/epos/ (accessed on 4 October 2021)

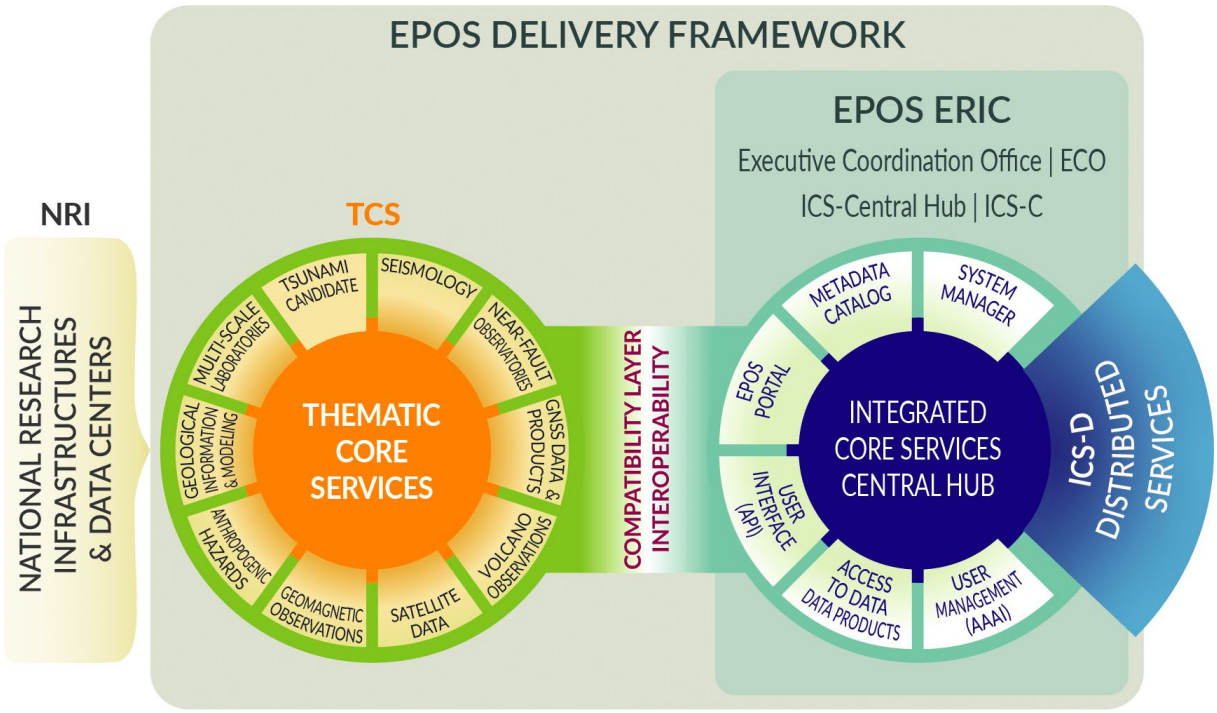

Figure 1. The Epos Delivery Framework includes – from left to right – National Research Infrastructures (NRI) and Data centres, 10 different Thematic Communities, and the Integrated Core Services. The latter is made up of a Central Hub for data integration, and distributed services for advanced analysis and visualization functionalities. These components constitute the EPOS Delivery Framework perimeter. EPOS Delivery Framework also includes the Executive and Coordination Office (Headquarter of the EPOS-ERIC legal seat) and the ICS-C Central Hub (technical hosting node). These two latter components constitute the new infrastructure coordinated by EPOS-ERIC through a defined governance structure and legal collaborative agreements. As such, they are part of the so-called EPOS-ERIC perimeter.

services, also further analysis and visualization. We focus on the FAIR principles as dealt with in the EPOS context, describing the FAIR data management practices in EPOS and EPOS synergies with other FAIR related initiatives. Finally, future directions are outlined on the basis of the experience gained in a decade with the creation of such a multidisciplinary distributed FAIR infrastructure.

2. Thematic Communities and multidisciplinary data integration

The assets provided by EPOS data providers include data, data products, software, services (DDSS) as well as sensor networks, laboratory equipment and associated Trans-National Access (TNA) programs, and computing facilities. These span over the entire Europe and can therefore be considered interconnected and part of a true international community.

The technical architecture to achieve such integration consists of four main elements, as shown in Figure 2: *Thematic Core Services, Integrated Core Services – Central Hub, Integrated Core Services – Distributed* and *Users*.

Thematic Core Services (TCS) (yellow block, left side in Figure 2) represent community nodes that provide access to DDSS from a specific domain at European level, and are fed by National RIs or International Organizations. TCS are the place where data and metadata harmonization occurs at the domain level, according to standards and practices adopted by the communities.

Integrated Core Services – Central Hub (ICS-C) (blue block, centre of Figure 2) is the place where true integration of heterogeneous DDSS occurs, and where scientists can access a one-stop shop where open multi-disciplinary assets are available. ICS-C represents the innovative aspect of EPOS and enables scientists to perform multi-disciplinary research in an easy way.

Integrated Core Services – Distributed (ICS-D) (light blue block, right side in Figure 2): these represent external service providers that make available added-value services like computational resources, storage or visualization and modelling tools, that can be integrated into the existing EPOS data portal. In this way, EPOS aims at maximizing the reuse of existing tools and facilities for analysing and processing data. Outside the computational and storage resources that are common to all domains, indeed, the ICS-D include discipline-specific tools that are currently made available to the scientific community in a scattered and standalone way.

Users (bottom green blocks in Figure 2) are represented by 4 main target groups, each of which has different requirements with respect to communication and data access. For instance, a scientist might be interested in specific fine-grained features of a dataset, in order to unravel new aspects of a physical phenomenon; a policymaker might be interested in aggregated information (e.g., bulletin or report) enabling him or her to make informed decisions; similarly, the private sector and society have specific needs that need to be addressd.

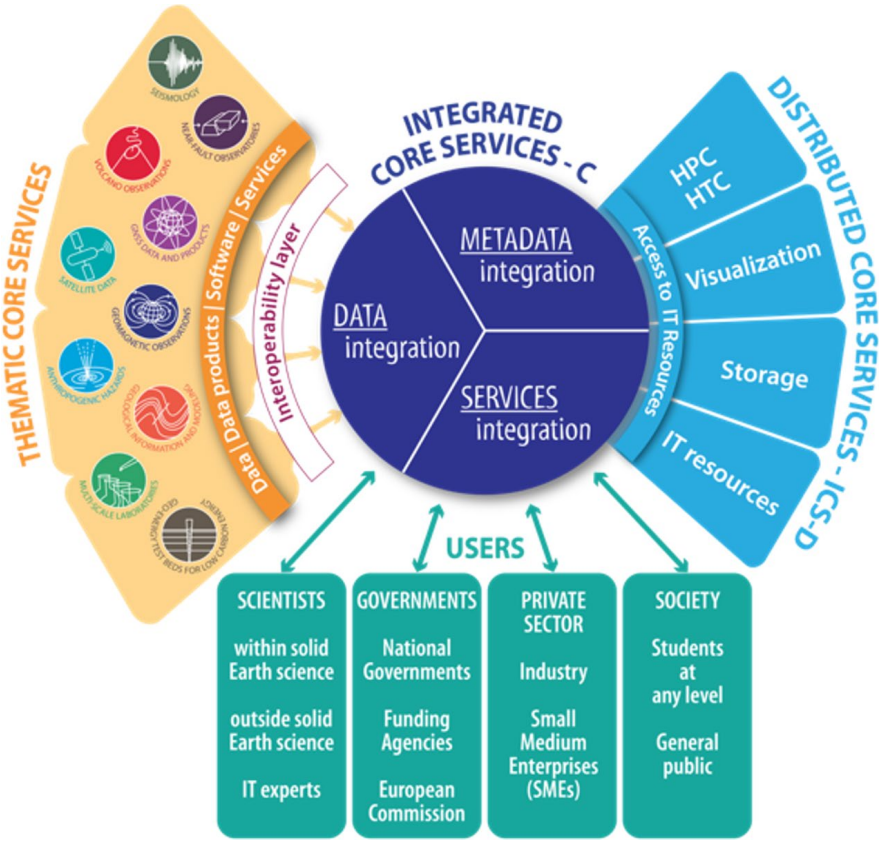

Figure 2. EPOS Architecture: the diagram describes the main components of the EPOS architecture and their relationships. The Thematic Core Services (TCS) (yellow, left side) harmonize the data provision at a European level guaranteeing that common standards and approach are used for the same type of assets. The ICS-Central Hub (blue, center) implements DDSS integration by means of an approach based on Data, Metadata and Services integration. The ICS-Distributed (light blue, right side) represent the access to external distributed IT resources that complement the data integration and access with advanced functionalities; these may include HPC/HTC facilities, visualization and storage services, other resources. The User component (green, bottom) is intended to represent users as key players also in the architecture since its design phase; they are categorized in four different categories (described in the article main text) and are those who – at the same time – raise the requirements and access to the provided services.

Data and data products integrated by EPOS Integrated Core Services are heterogeneous in terms of their representations: these include a) georeferenced data, which identify a geographic location and characteristics of the Earth's surface. They are typically recorded in terms of latitude and longitude or by using some form of Cartesian coordinate system (e.g., maps or geo-located equipment like seismic stations), b) time series, here defined as set of

regular time-ordered observations of a quantitative characteristic of an individual or collective phenomenon, taken at successive periods of time measured by a detector (e.g., GPS station, geomagnetic stations), c) non-georeferenced data, which cannot be specifically traced to a specific location or area (e.g. software packages or pdf reports).

Table 1 provides an overview of service types grouped by communities (TCS), and emphasizes the heterogeneity in the adoption of standard specifications and data representations. More detailed information about specifications, formats and standards adopted by the communities are reported in Appendix A.

Such heterogeneity presents challenges that require new approaches in terms of methodologies, technical design and community building, as described below.

TCS Name	Service Group Name	# Services	Service standard	Data Representations
Anthropogenic Hazards				
Anthropogenic Hazards	Anthropogenic services	2	Custom	georeferenced, non-georeferenced
Anthropogenic Hazards	Anthropogenic episodes	37	Custom	georeferenced, non-georeferenced
Geological Information and Modeling				
Geological Information and Modeling	Borehole data	2	OGC[4]	georeferenced
Geological Information and Modeling	Geological maps	2	OGC	georeferenced
Geological Information and Modeling	3D/4D models	2	OGC	georeferenced
Geological Information and Modeling	Mineral resources	2	OGC	georeferenced
Geomagnetic Observations				
Geomagnetic Observations	Geomagnetic data	7	Custom	georeferenced, time-series
Geomagnetic Observations	Geomagnetic models	2	Custom	non-georeferenced
Geomagnetic Observations	Geomagnetic indices and events	6	Custom	georeferenced, time-series
Geomagnetic Observations	Magnetotelluric models and data	4	Custom	georeferenced, non-georeferenced, time-series
GNSS Data and Products				
GNSS Data and Products	GNSS data	3	Custom	georeferenced, non-georeferenced
GNSS Data and Products	GNSS data products	11	Custom	georeferenced, time-series

4 Open Geospatial Consortium: https://www.ogc.org/standards (accessed on the 17th of January 2022). In this context we refer to the *family* of standard, as the service may have adopted more than one for the same data product (e.g., Web Map Service – WMS, and Web Feature Service – WFS).

TCS Name	Service Group Name	# Services	Service standard	Data Representations
Near Fault Observations				
Near Fault Observations	Seismological data	26	FDSN[5], Custom	georeferenced, time-series
Near Fault Observations	Geochemical data	7	Custom	georeferenced, time-series
Near Fault Observations	Geophysical data	4	FDSN	georeferenced, time-series
Satellite Data				
Satellite Data	InSAR data	8	Open Search[6]	georeferenced, non-georeferenced
Seismology				
Seismology	Waveform services	42	FDSN, Custom	georeferenced, non-georeferenced, time-series
Seismology	Seismological products	11	FDSN, OGC, Custom	georeferenced, non-georeferenced
Seismology	Earthquake hazard and risk products	6	OGC	georeferenced
Volcano Observations				
Volcano Observations	Seismological data	4	FDSN	georeferenced, time-series
Volcano Observations	Geodetic data	7	FDSN, Custom	georeferenced, time-series
Volcano Observations	Geochemical data	4	Custom	georeferenced, non-georeferenced
Volcano Observations	Satellite data	5	Custom	georeferenced
Volcano Observations	Ground based remote sensing data	5	Custom	georeferenced
Volcano Observations	Volcanological/ petrological data	9	Custom	non-georeferenced
Volcano Observations	Geohazards products	10	Custom	georeferenced, non-georeferenced

Table 1. An overview of service types grouped by communities (TCS) and the heterogeneity in the adoption of standard specifications and data representations.

5 The International Federation of Digital Seismograph Networks (FDSN): https://www.fdsn.org/webservices/ (accessed on the 17th of January 2022)

6 Open Search standard: https://opensearch.org/docs/latest/ (accessed on 17th of January 2022)

Data and data products are produced by a heterogeneous set of data providers, providing data with various levels of maturity that can be categorized following the data taxonomy given in table 2.

Level Code	Data Processing Level
Level 0	Raw data, or basic data
Level 1	Data products coming from nearly automated procedures
Level 2	Data products resulting from scientists' investigations
Level 3	Integrated data products coming from complex analyses or community shared products
Level 4	Software, IT tools

Table 2. Data taxonomy and processing levels.

In the EPOS architecture, data are usually collected/created at national observational systems (monitoring networks). As a consequence, in most cases data are stored locally and maintained (archived, curated and managed) by data providers. While this is the case for most of the level-0 (raw data) and level-1 (automatically processed) data, there is a variety of data products (level-2) that are created at later stages of the data life cycle (Figure 3), following a set of operations such as pre-processing, adopting quality assurance, conducting further processing,

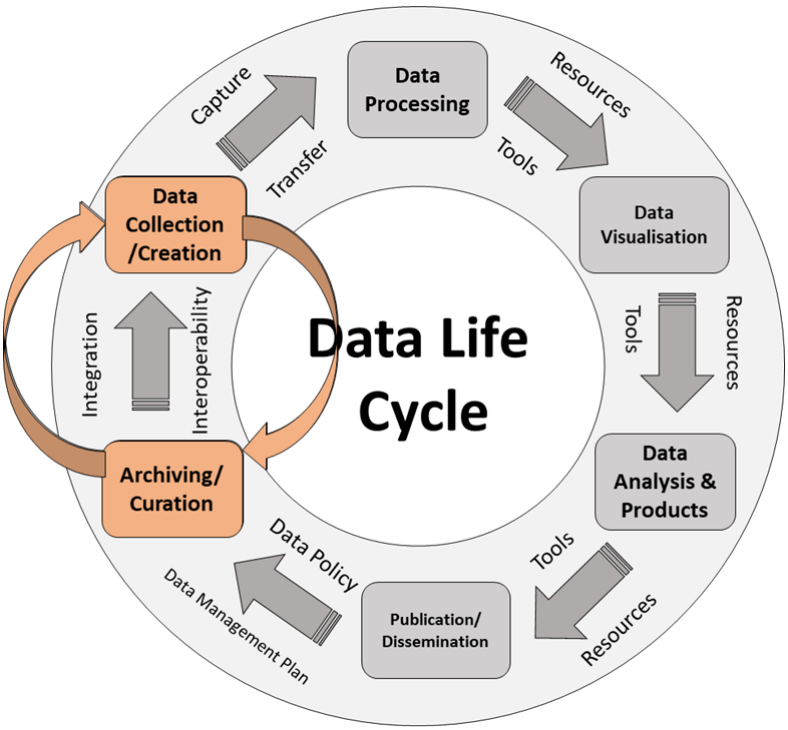

Figure 3. The diagram shows a typical data life cycle in EPOS, which involves either data collected by individual research infrastructures (RIs) through distributed sensor networks or through higher-level data products created at individual RIs.

visualizing and analysing before a final data product is created. In addition, higher levels of data products (level-3 and level-4) are provided from scientific communities based on complex interactions between various subdisciplines of solid Earth science through advanced integration (e.g., modelling and simulations) of multidisciplinary data. Earthquake hazard and risk maps [Giardini et al., 2014], anthropogenic hazard episodes [Orlecka-Sikora et al., 2020], simulations of volcanic eruptions, earthquake wave propagation within the Earth structure and tsunami wave propagation in the oceans are such examples.

In Figure 3 the main elements of the data lifecycle of the EPOS delivery framework are shown. Several models exist in literature [Sinaeepourfard et al., 2015], for describing the data lifecycle from production stage to publication stage in a given landscape. The EPOS data lifecycle was created by leveraging the DataOne [Michener et al., 2012] and Digital Curation Centre (DCC) models [Sinaeepourfard et al., 2015]. The EPOS data lifecycle covers all levels of data, as defined in the above data taxonomy, in a twofold way: on one hand, data from the various levels are all integrated and made interoperable in the "Integrated Core Service-Central Hub (ICS-C)"; on the other hand, through the EPOS Data portal and the underpinning e-Infrastructure, EPOS can manage the data lifecycle in its entirety (except publication) by easing the access to or production of data at any level.

Archiving and curation of data or higher-level data products are usually handled by the data providing institutions (RIs). It is important to note that various data providing institutions in EPOS follow their relevant community standards for data acquisition, storage, archiving and curation. A wide range of data and products are included in EPOS stemming from the different stages in the data life cycle. Managing such heterogeneous data, therefore, requires clear data management plans in place at both individual RI level, as well as the thematic core services (TCS) and integrated core services (ICS). Integration of new data products with already archived data and allowing open access to interoperable data involves data management in several dimensions such as technical, governance, legal, financial, strategic, policy, resources, security, privacy, sensitivity, ethical, data quality, metadata and provenance.

3. EPOS approach to Data Integration

With more than 220 different types of services integrated, EPOS is a prominent example of Information-Powered Collaborations (IPC), which are defined as "complex, dynamic and heterogeneous environments that enable information sharing among actors (e.g., researchers, scientists, practitioners, agents) from independently managed organizations (e.g., research institutes, resource providers), thereby supporting knowledge and expertise exchange in a multidisciplinary context. […]" [Trani et al., 2018].

Systematic approaches are required to exploit the full power of such collaborations. For instance, structured methodologies can help establish focused interactions that address specific aspects (e.g., concepts, representations and instances) within a broad and complex context [Trani, 2019].

Here we focus on the technical implications of established agreements where interoperability, as defined by the FAIR principles [Wilkinson et al., 2016], plays a crucial role. It supports homogeneous, harmonized views and a common understanding of heterogeneous resources and assets. It offers usable data, tools and an agreed way of working to better employ them. It provides users and stakeholders with the right amount of information for their intended applications.

The FAIR principles can support the shaping of the technical backbone underpinning an IPC by targeting key requirements of findability, accessibility, interoperability and reusability. The principles propose technological solutions (e.g., Persistent Identifiers (PIDs), metadata, catalogues, Authentication and Authorization Infrastructure (AAI) systems). However, they leave room for specific implementations. They can serve as guidelines for promoting discussions and interactions, thus allowing their application within established communities and practices. In the following sections we discuss the EPOS approach for the application of the FAIR principles.

Achieving integration of heterogeneous data sources from the different communities within EPOS IPC communities using different standards for their assets (e.g., datasets, services), and for the metadata describing them, is a huge challenge that cannot be tackled at a technical level only. Indeed, challenges and efforts to build an IPC, such as EPOS, cover several aspects e.g., scientific, governance, sustainability, financial, collaborative, policy, organizational, scientific and legal dimensions, as they are addressed in other publications of this special issue.

In this context, it is worth mentioning the example of policies, a vital element for achieving data integration, to emphasize the strong relations among different dimensions: policies, together with the guidelines (enactment of policies for business purposes) and the implementation (technology support) may help or hinder interoperation.

For example, a policy may preclude making certain assets available (hindering) or a policy may demand that all assets have the same metadata standard (helping).

Drilling down to the more technical aspects of data integration in EPOS, two main elements were recognized as necessary for data integration: the *metadata* and the *service-based architecture*.

As for the metadata, it has to cover both syntax (structure) and semantics (meaning) of data values in the metadata describing assets and in the assets themselves. The syntax needs to be as rich or complex as necessary to accurately represent the real world, as also prescribed by the FAIR principles. Semantics require a similar structure so that a term in one vocabulary can be related not only to other terms in that vocabulary but to terms in other vocabularies, including multi-linguality. In the context of EPOS, a metadata catalogue addressing these requirements is used: CERIF[7] (Common European Research Information Format; an EU Recommendation to Member States) [Jeffery et al., 2014].

As for the system architecture, it is required to manage the complex landscape detailed by the metadata and to satisfy user requests. In the case of EPOS, the system design is inspired by the Microservice approach [Newman, 2015; Richardson, 2017], where each microservice implements a clear function. In particular, converter microservices ensure that mappings are done from many different metadata formats to a canonical format for – for example – visualizing the data in an integrated way in the EPOS Data portal, as described in the next section. In addition, this approach relies on the usage of web services by the data providers as the mechanism for making data, metadata and other assets accessible.

In the EPOS experience, considering only one of two elements alone, i.e., only metadata or only system architecture, is not sufficient for building a Research Infrastructure for data integration.

Interestingly, the FAIR principles do not explicitly mention web-service-based systems, delegating to the data practitioners and engineers the design and implementation of the data stewardship nodes. This is also discussed in [Koers et al., 2020], where, following the inputs from various stakeholders' groups in the context of three workshops organized by the projects FAIRsFAIR, Research Data Alliance (RDA) Europe, OpenAIRE, EOSC-hub, and FREYA, a set of guidelines was established, and actions suggested. For the Service providers stakeholder group, the authors suggest making repositories support FAIR by developing tools, such as APIs, and share best practices and user stories. However, the establishment of a clear, robust and systematic approach for FAIR system development, would further improve the understanding of the technical implications of FAIR principles and ease the design and implementation of FAIR data stewardship systems, as already discussed in another work [Bailo et al., 2020].

Once metadata and system challenges are tackled, different approaches can be pursued with respect to data and metadata interoperation.

The brokering approach, for example in GEOSS[8] [Nativi et al., 2014] uses software hard-coded to convert metadata or digital assets from one standard to another. Such software is difficult to maintain because any change of format used for metadata or an asset requires re-coding of the converter broker. Worse, since brokers work pairwise, if there are *n* metadata standards or asset standards to be converted, *n(n-1)* broker converters are required.

The metadata catalogue approach, as used in libraries[9], in contrast to the brokering approach, reduces this to *n* convertors to/from the superset canonical catalogue metadata schema. This is a huge reduction of effort, and the mappings used for metadata conversion can be re-used for asset conversion.

The web-services approach, usually adopted in pure microservices architectural designs [Richardson, 2017], assumes that any data source uses web services to interoperate and that the integration is done by the consumer, usually a Graphic User Interface. It provides huge flexibility but can overload the data integration GUI when the required integration is complex (e.g., geo-referenced data combined with time-series data).

In the case of EPOS, none of the above approaches adopted alone served the purpose of data integration. In fact, EPOS has chosen a mixed approach, combining the most advantageous aspects of metadata and services with brokering [Nativi et al., 2015]. This means that brokers convert from many metadata formats, exposed to the consumers as web services, to one central system (ICS-C) using a rich metadata catalogue, thus providing homogeneous metadata descriptions and management of assets.

7 https://eurocris.org/services/main-features-cerif (accessed on the 4th of October 2021)

8 https://www.geoportal.org/about (accessed on the 4th of October 2021)

9 https://www.worldcat.org/ (accessed on the 4th of October 2021)

4. The EPOS Data Portal

The EPOS Data Portal combines metadata and service-based architectural approaches mentioned above by relying on a multi-tier architecture [Schuldt, 2009], which enables the flexibility for enhancing the system according to the evolutionary requirements. The reliability of the tier components, that can be restored in case of disruption, without interfering with each other, allows shorter system recovery times. Further, the multi-tier approach simplifies development as different teams may work on each tier at same time without interfering in each other's tasks or domains.

The architecture, depicted in Figure 4, consists of four tiers, from top to bottom: Graphical User Interface (GUI), Integrated Core Services, including a Central Hub system (ICS-C) and Distributed systems (ICS-D), Interoperability tier and Thematic Core Services (TCS).

For the ease of readability, ICS-C and ICS-D are discussed in different sections.

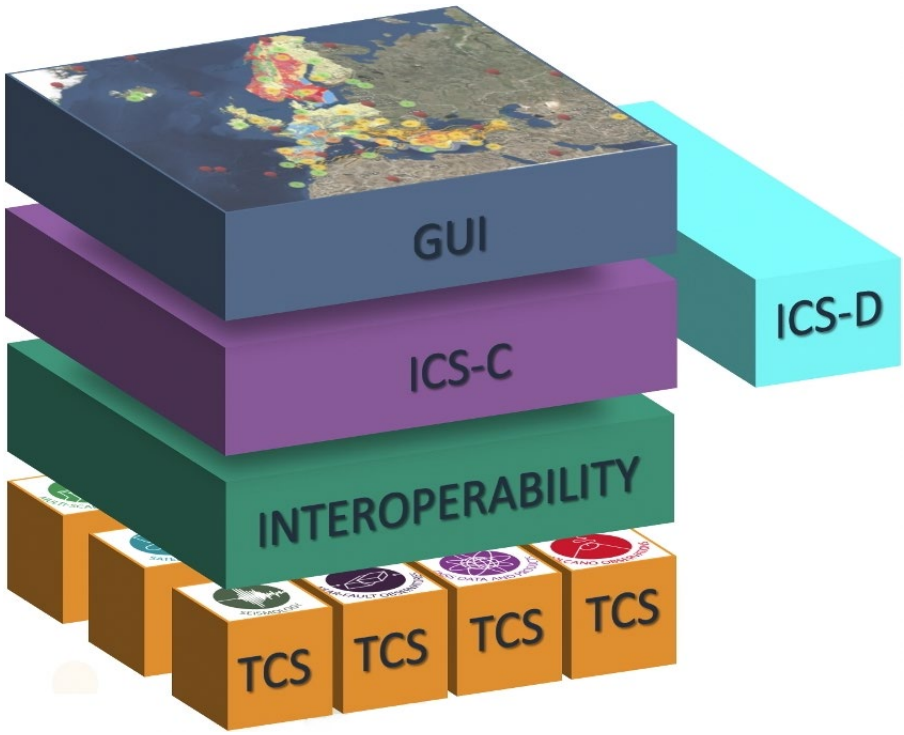

Figure 4. The EPOS Data Portal Architecture consists of 4 tiers, Graphic User Interface (GUI), Integrated Core Services (Central Hub and Distributed), Interoperability, Thematic Core Services (TCS).

4.1 Graphical User Interface

The Graphical User Interface (GUI) provides access to DDSS provided by different TCS.

The EPOS GUI[10] consists of: (i) a search area (Figure 5a) which enables users to filter data by using several criteria (e.g. spatio-temporal extents, keywords, data/service providers, free-text); (ii) a data pre-visualization area (Figure 5b) used to pre-visualize the selected data on Map, Table or Graph (even in overlap mode); (iii) a metadata and sub-setting area (Figure 5c) intended to provide details about the selected data (e.g., name, description, license, DOI), as well as to further refine the search in order to dig into a smaller level of granularity of data. Favorite data can be downloaded in different formats and can be also added to a user workspace for further analysis and advanced processing.

10 https://www.ics-c.epos-eu.org/ (accessed on 20th of October 2021)

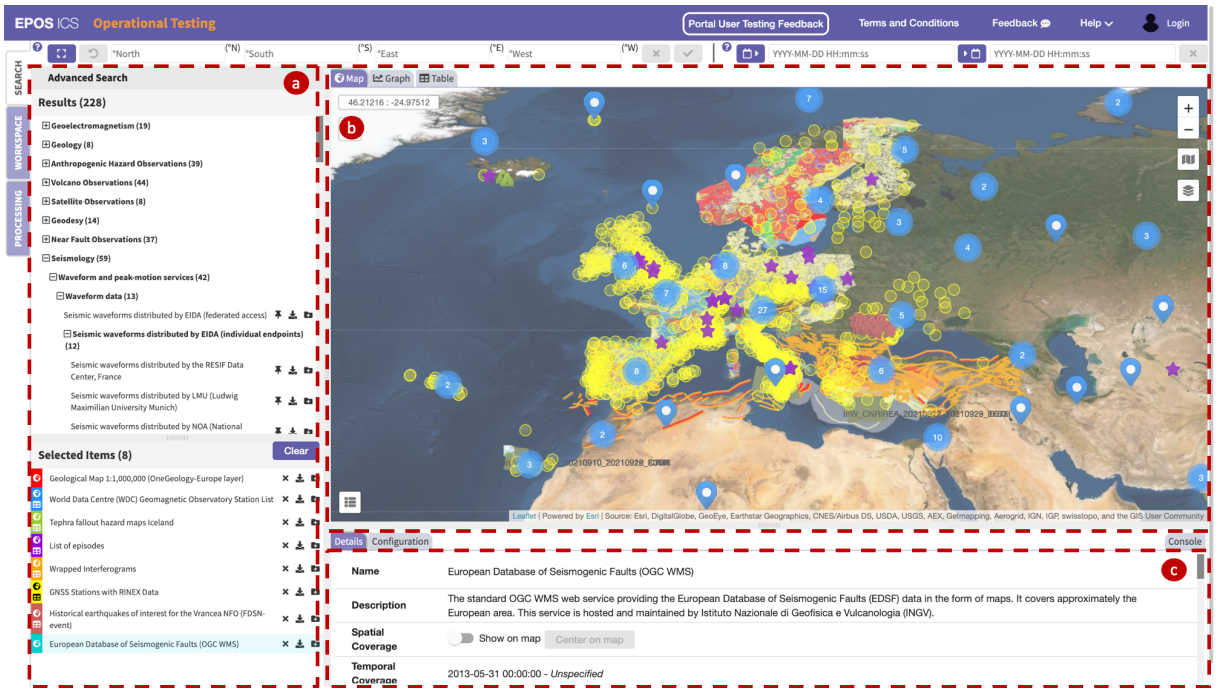

Figure 5. The EPOS Data portal Graphical User Interface, consisting of: search area (a), data pre-visualization area (b), metadata and sub-setting area (c).

4.2 Integrated Core Services – Central Hub system (ICS-C)

The Integrated Core Services – Central Hub system (ICS-C) is a component of the ICS tier, i.e. the place where multidisciplinary resources are integrated.

It consists of three sub tiers:

1) *WebAPIs*, which provide a set of RESTful [Richardson et al., 2007] endpoints to enable the communication with GUI tier. These endpoints are used by the GUI for data discovery, data access and retrieval, as well as managing user's workspace.
2) *Software components*, which are developed according to the microservices paradigm [Newman, 2015]. They enable ICS to run in a distributed environment, to adopt a software-independent approach ensuring up-to-date technological upgrades, to properly scale specific system functionalities, to enhance reliability, to isolate independent software applications running in a shared environment.
3) *Metadata Catalogue*, which is used to store information about data, data products, software, services, and other information associated with them. The catalogue is based on CERIF (Common European Research Information Format) data model [Bailo et al., 2014] which is able to represent the real-world entities and attributes of interest to EPOS.

4.3 Interoperability

The Interoperability tier enables the interaction between ICS and TCS by defining a common knowledge representation language. An extension of DCAT-AP[11], namely EPOS-DCAT-AP[12], has been developed and adopted to describe the diversity and heterogeneity of TCS assets. EPOS-DCAT-AP is represented in RDF/Turtle format. The metadata collection of EPOS-DCAT turtle files by TCS is carried out in stages by prioritizing specific metadata en-

11 https://joinup.ec.europa.eu/collection/semantic-interoperability-community-semic/solution/dcat-application-profile-data-portals-europe (accessed on the 4th of October 2021)

12 https://github.com/epos-eu/EPOS-DCAT-AP (accessed on the 4th of October 2021)

ties. Collaborative tools (e.g., GitLab[13], GitHub[14], EPOS Metadata Editor[15]) have been provided to TCS in order to support and sustain the collection process and hence to populate the metadata catalogue. With this approach, each community can keep using its own specific metadata standard, delegating to EPOS-DCAT-AP only the description of the services delivering the data (or data products). EPOS-DCAT-AP is converted to CERIF for the catalog.

4.4 Thematic Core Services (TCS)

Thematic Core Services (TCS) represent datasets and services provided by domain specific communities. Their data types, formats, protocols, access methods and policies, are highly heterogeneous.

A harmonization activity has been promoted and stimulated at community level by fostering the creation of new European-wide thematic nodes and supporting existing organizations.

4.5 Integrated Core Services – Distributed (ICS-D)

The Integrated Core Services – Distributed (ICS-D) is a component of the ICS tier, designed with the specific aim of integrating into the ICS-C external services provided outside of the EPOS delivery framework: computational earth science applications, advanced visualization services, benchmarking services, training services and services from other environmental domains and others.

In the design of ICS-D, three essential elements, that are common to any of the categories above, were identified: (i) workflow engine (WE), (ii) virtual research environment (VRE), and (iii) computational resources. A workflow engine is needed in order to allow users to manage, compose and deploy their desired workflows for computations as well as for processing. Virtual research environment (VRE) – interpreted in one of its main flavours [Candela et al., 2013] is another essential element needed to create a sand-box (a containerized environment) for researchers to bring together elements that are needed for their individual research. Computational resources are needed primarily for computational Earth science (CES) applications demanding access to supercomputing facilities providing high performance computing (HPC) or high throughput computing (HTC). How to ease the access to such infrastructures, providing general-purpose services and tools to users to perform complex analysis, is a challenge that the EPOS team have approached in several R&D activities, in cooperation with the seismology TCS [Atkinson et al, 2019], [Klampanos et al, 2019]. The users of the EPOS community should be able to discover the availability and fitness for purpose of a particular ICS-D via the interrogation of the ICS-C catalogue. Here metadata describe the ICS-D details like access endpoint, capabilities and responsible contacts, thus fostering FAIRness of operations, as well as traceability of the products generated by the users via these services.

4.6 EPOS Workspaces and ICS-D prototype

Through the collaborative work with the ENVRI-FAIR project [Petzold et al., 2019], the prototype of a Web API that delivers a collection of integrated and general-purpose ICS-Ds was developed. The SWIRRL-API [Spinuso et al., 2020; Spinuso et al., 2021], is designed to enable community portals, such as the EPOS Data Portal, to build interactive and provenance-aware workspaces. Thanks to SWIRLL, EPOS empowers its users to conduct investigations and develop analyses with the assets selected by means of the EPOS portal discovery service. A prototype application was developed, in which the EPOS Data Portal interacts with the API to assemble the workspace as a comprehensive VRE that provides researchers with programming and visual analytics tools, such as JupyterLab[16] and Enlighten-web[17] [Langeland et al., 2019], respectively. The API hides the complexity of deploying and orches-

13 https://about.gitlab.com/ (accessed on the 4th of October 2021)

14 https://github.com/ (accessed on the 4th of October 2021)

15 https://epos-eu.github.io/SHAPEness-Metadata-Editor/ (accessed on the 4th of October 2021)

16 https://jupyter.org/ (accessed on the 4th of October 2021)

17 https://demonstrator.webfarm.cmr.no/covid19/doc/enlweb/html/index.html (accessed on 20th of October 2021)

trating the resources needed by the VRE on a hosting cloud e-infrastructure that exposes a Kubernetes cluster[18]. The data is staged to the VRE via the execution of a workflow, which makes sure that the data is shared between the different tools, so that these can be used to perform analysis in combination. Workflows can be several and are developed independently by professional research-developer and engineers. Once integrated in SWIRRL, they can be controlled via the API that makes them available to the portal. Thus, depending on their implementations and the user's choice, the workflows can perform common operations, such as data reduction, pre-processing, subsetting etc. Finally, SWIRRL generates and stores metadata-rich provenance of the operations performed in the VRE. This is used to support the traceability, recovery and reproducibility of the analyses and their supporting environment. This is a fundamental capability of the prototype, with the objective of delivering FAIR analysis services by design.

5. FAIR Data Management

Firstly developed in the FORCE-11 group [Wilkinson, 2016; Mons, 2017; Collins et al., 2018], the FAIR principles have become *de-facto* reference criteria for many European Open Science Cloud related initiatives (e.g. ENVRI-FAIR and EOSC-Life in the past H2020-INFRAEOSC-4-2018[19] call) and EU calls for proposals (e.g., "Research infrastructure services to support health research, accelerate the green and digital transformation, and advance frontier knowledge (2021)" (HORIZON-INFRA-2021-SERV-01)[20]) where they are listed as cross-cutting Priorities. Lately, their area of competency is expanding to services [Koers et al., 2020] and to policies, meaning that especially in the Earth-science field, editors will insist that key data are made available in repositories that support the FAIR principles [Stall et al., 2019].

In this context, a burning question arises: how is EPOS positioned with respect to FAIR principles?

5.1 EPOS approach to FAIR

EPOS being such a wide community, there are of course several shades of adoption and compliance to the FAIR principles, but when it comes to the IT design and developments activities, and in particular the Integrated Core Services, we can firmly state that EPOS is not simply endorsing FAIR principles, but it was – to some extent – foreseeing the technical implications of FAIR.

Indeed, the EPOS technical Work Package leader in EPOS-PP participated in FORCE-11[21], and the architecture underpinning the data portal has considered *metadata* and *service-based architecture* as the main driving concepts since its early stages [Jeffery et al., 2014]. The emphasis on the formal syntax and declared semantics [Bailo et al., 2015] as requirements for a rich metadata model ensures that most of the FAIR requirements related to Findability and Accessibility are satisfied, as described below. In addition, individuals with key roles in the design of the EPOS IT Architecture, are members of RDA[22] working groups and Interest groups (e.g., Metadata Interest Group and Virtual Research Environment Interest Group).

The above demonstrates a deep and wide involvement of EPOS IT leaders in the European and international FAIR activities.

On the purely technical dimension, EPOS IT Team has developed a well-defined roadmap for compliance with the FAIR principles. FAIR are indeed "principles", meaning that they are abstract in nature and can be applied in different ways. In order to implement these in the specific EPOS agile development team context, a methodology was needed. It was developed and is described in [Bailo et al., 2020]. On the basis of experience and know-how in the EPOS community of practitioners, experts, and engineers, a common approach was observed, which is described by the re-organization of FAIR principles into a four-stage roadmap that considers the four conceptual lay-

18 https://kubernetes.io/ (accessed on the 4th of October 2021)

19 https://cordis.europa.eu/programme/id/H2020_INFRAEOSC-04-2018 (accessed on the 12th of January 2022)

20 https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/opportunities/topic-details/horizon-infra-2021-serv-01-01 (accessed on the 4th of October 2021)

21 https://www.force11.org/node/5003 (accessed on the 4th of October 2021)

22 https://www.rd-alliance.org/ (accessed on the 20th of October 2021)

ers guiding the architecture design and implementation of EPOS RI, that is to say: a) Data, b) Metadata, c) Access, d) Use (of services and data) (Figure 6).

Data, which is the core element and the starting point of the EPOS data life cycle, is the first layer. Once data is defined, the next step is usually the creation of metadata for discovery and contextualization, possibly according to existing standards in a specific domain. Once these two aspects (data and metadata) are defined the real data provision needs services for data access (third stage, including for instance standard based web-services) and services for data usage (fourth stage, including for instance processing APIs). According to such a four-stages pyramid, two elements were then defined: i) the FAIR principles that need to be considered at each stage, and ii) the types of technologies required to satisfy the FAIR principles at a certain stage.

The definition of such a roadmap supports the IT staff in the domain community to approach and solve practical problems like how to properly describe and store data and metadata, or how to provide authenticated access to datasets according to a set of predefined policies. The natural evolution of this approach could be the provision of a reference architecture.

For each of these layers, principles to be addressed were defined, and technologies for their adoption were selected, as shown in Figure 6.

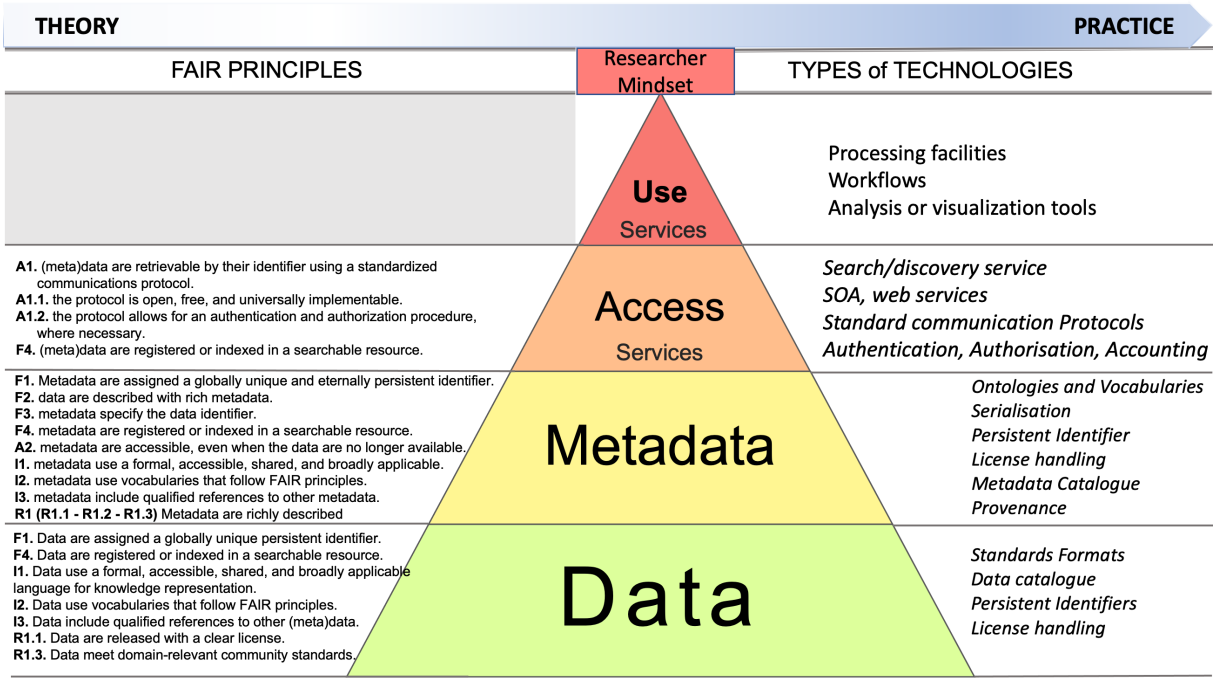

Figure 6. Four stages conceptual approach. The pyramid in the centre represents an approach that reflects the mindset and the practices of technical staff in the data provider nodes, whose native task is to produce data for science.

For instance, in order to comply with Accessibility requirements at the Access level (A1, A1.1, A1.2) web services for accessing the entire infrastructure, including the metadata catalogue, were developed, and standards for Authentication adopted (e.g., Oauth2[23]).

In the case of Reusability, instead of undertaking specific technical activities in one or more of the four layers described above, other types of actions were considered. Indeed, whenever a (meta)data policy is required (e.g. R1.1), a community effort to discuss and adopt specific licenses was pursued. Due to the width of EPOS community, the delicate matter of policies, and hence licenses, has been approached in an incremental way and has required a long time. A first version of policies was firstly discussed in the context of EPOS Preparatory Phase project[24] and

23 https://oauth.net/2/ (accessed on 4 October 2021)

24 EPOS-PP (Preparatory Phase, 2010-2014), where about 20 partners joined the project

then released as EPOS Data Policy[25] in the EPOS Implementation Phase project[26]. The latter document provided guidelines about licenses on data and metadata: for instance, "To ensure the widest dissemination and publicity for EPOS managed DDSS, it is essential that metadata are easily and freely accessible at any time, with as few restrictions as possible. In order to achieve this, Suppliers will be encouraged to affix open licenses, preferably Creative Commons 4.0 CC:BY[27], to their metadata."[24]. Currently, policies are being re-discussed in the context of EPOS-ERIC, where a clear governance including agreements to rule relations between EPOS-ERIC and TCS communities are established, as described in the governance sub-section below.

5.2 EPOS challenges for FAIR technical implementation

The description of all the best practices and technical approaches to comply with the FAIR principles is out of the scope of the current work as these are discussed in [Bailo et al., 2020]. However, it is worth mentioning that in order to make the EPOS IT infrastructure compliant with FAIR in a sustainable way, the design, implementation and validation activities had to tackle interesting challenges in the following areas: a) *metadata*, where the main challenge is to be able to map all the relevant information about data and assets by TCS data providers; metadata (with formal syntax) enables the system to make autonomic decisions and enables users to access all the information they need to contextualize data and services they are searching for; b) *semantics*, where the high heterogeneity of the communities in terms of metadata formats forced the IT Team to adopt flexible solutions and to plan mappings among different metadata standards, each with its own semantic representation; c) *interoperability*, that requires on one hand the adoption of widely used standards for communication and (meta)data provision, and on the other the development of a micro-service based architecture that ensure maximum flexibility and extensibility of the system; d) *Authentication and Authorization*, where the adoption of AARC blueprint architecture [AARC Community members et al., 2019] was fundamental to ensure Accessibility and Findability requirements; *e) community building*, a challenge often neglected and further discussed in the next section, that stimulated the creation of an agile methodology inspired by the shape-up method[28] in order to manage developments and interactions with the wide EPOS community of IT practitioners, both from the main Central Hub (ICS-C) and from the various thematic communities (TCS).

It is worth mentioning, in this context, that the non-prescriptive [Mons et al., 2017] nature of FAIR principles leaves, on one hand, a good extent of freedom to Ris for what concerns the implementation of technical solutions, but on the other hand it puts a huge burden on the communities that are forced to find solutions for complying with FAIR principles. At the same time the freedom of implementation becomes a forge of new ideas, solutions, methodologies and innovative potential that deserves to be supported and taken into consideration when planning wide-ranging actions at International and European level.

5.3 EPOS Governance for supporting FAIR

In EPOS, in order to guarantee the sustainability of the technical activities required to maintain a FAIR data stewardship system for data integration, available through the EPOS Data Portal, a clear FAIR governance was established: as a natural consequence of the EPOS architecture, FAIR data management is primarily handled by the national research infrastructures (NRIs), which are usually responsible for collecting, storing, archiving, curating and providing access to data. While finding and accessing is dealt with by the data provision at NRIs, interoperability requires integration of data provided by several NRIs through thematic services. Thematic Core Services (TCS), that constitute the second layer in the EPOS architecture, deal with the data management and governance through legally binding collaboration agreements signed between the TCS consortia and the EPOS-ERIC. As such, a high level of harmonization and standardization work is done at TCS level providing interoperability within the

25 https://www.epos-eu.org/sites/default/files/2020-12/EPOS%20DATA%20POLICY_July2018.pdf (accessed 18th of January 2022)

26 EPOS-IP – EU Horison2020 – InfraDev Programme – Project no. 676564; 2015-2019, included 47 partners plus 6 associate partners from 25 countries from all over Europe and several international organizations

27 https://creativecommons.org/licenses/by/4.0/ (accessed 18th of January 2022)

28 https://basecamp.com/shapeup (accessed on the 4th of October 2021)

domain specific data in individual TCSs. The real interoperability, in the true sense of making multidisciplinary data available in an integrated way from various thematic communities, is handled at the Integrated Core Services (ICS) level, where multidisciplinary data are harmonized and made accessible through the EPOS data portal.

In the EPOS governance structure, further discussed in another work in this special issue, the interactions between the ICS and TCS, as well as developments and operation of the entire ICS system is coordinated by a dedicated IT-Board which consists of representatives from the thematic communities appointed by the Service Coordination Committee (SCC), from hosting organizations responsible for the operation of the system and representatives from other EPOS-ERIC bodies, including the IT-Unit, ICS-TCS coordination, Scientific officer etc., and is chaired by the IT-officer of EPOS-ERIC. The IT-Board reports to the Executive Director and the Executive Committee ensure that important decisions are consulted in the relevant bodies before their approval in the General Assembly.

Through the IT-Board and the underlying mechanisms that are in place, together with clear links to the governing and decision bodies in EPOS-ERIC, sustainability of the entire ICS system is guaranteed. Necessary resources are secured by the individual (legally binding) multi-year collaboration agreements (MYCA), that are signed between various organizations, regarding TCS governance and coordination, ICS-TCS interactions, ICS-C hosting and operation and other potential key players.

5.4 EPOS community challenges for supporting FAIR

EPOS community is huge, being composed by more than 250 individual research infrastructures or institutions, organized, as said, in 10 thematic communities. This poses interesting challenges when it comes to cross-community technical work where a wide number, between 70 and 100 i, of technical professionals are involved.

On the basis of the requirements raised by the EPOS strategic plan and the EPOS Thematic communities, technical objectives are set and prioritized annually by the IT-Board. Activities for achieving the objectives are executed by relevant groups of developers, through well-structured cyclic community workshops bringing together the IT-developers and scientific experts from both thematic communities (TCS) as well as the ICS. These ICS-TCS interaction workshops are arranged four times a year, where in each workshop, development plans for the following three months are made and are executed in well-defined tasks (pitches) following the already mentioned “shape up” methodology. The shape-up method was developed for enabling a constant delivery while keeping the organizational overhead very light. What happened in EPOS before the adoption of this method was that the definition of tasks, deadlines and resources allocation on the basis of the requirements was done in a very detailed way, but when it came to deadlines, difficulties were encountered because each individual from different institution had its own priorities and was bound to local management practices.

The shape-up innovation is that deadlines are set in advance: the work is carried out by means of “pitches” each of which can be executed in a given amount of time (typically 8 weeks, extended to 10 weeks in the EPOS case). On this basis, each organization does its best to manage resources, and tasks are allocated based on the available time, thus reducing the risk of slipping deadlines or, even worse, not to ship at all.

With this approach, difficult challenges could be addressed. For instance, the adoption of EPOS-DCAT-AP as a standard to exchange knowledge between the diverse TCS and ICS was addressed in an incremental way through different pitches. Direct interactions with the communities have been set up by organising dedicated meetings and creating task forces that included domain and metadata experts. Collaborative tools, like a metadata editor [Paciello et al., 2021], have been provided to support and sustain those interactions. Preliminary mappings of a list of prioritised resources were produced and collected in the EPOS-DCAT-AP GitHub repository. Such an incremental process, based on well defined, time-bound pitches, helped refining the model and validating the metadata standard.

The above-described approaches and governance structure are necessary for operating and ensuring sustainability of the EPOS FAIR Data Portal, and as such it is appropriate to refer to it as FAIR governance. It corroborates the idea that FAIR principles, when implemented in Research Infrastructures that need to guarantee – by mandate – operations with a high availability level, cannot be limited to mere technical aspects. Their impact on other dimensions (governance, sustainability) should indeed be seriously considered by the FAIR international community. Also, it demonstrates that while technical viability of FAIR principles can be tested in specific contexts like competence centers or demonstrators in European initiatives, the real sustainability, benchmarking and actual realization of FAIR principles is usually done in the context of Research Infrastructures like EPOS, committed to provide access to data and services to the scientific community and beyond.

6. EPOS FAIRness in a Pan-European landscape

The research communities internationally, and particularly the IT and librarian communities concerned with data centres, have embraced FAIR as a concept and are putting it into practice, as evidenced by the report of one of the Working Groups of the EOSC Executive Board [Hong et al., 2020]. Several EC-funded projects are concerned with promoting and measuring FAIRness, as for instance GO-FAIR [Schultes et al., 2018], FAIRsFAIR[29], FAIRsharing .[Sansone et al., 2019] and many others. Other initiatives exist, more dedicated to managing FAIR in an RI context, where EPOS has strong linkages and influence. For instance, EPOS represents the solid earth in ENVRI (Environment RIs), a cluster of 26 environmental RIs [Petzold, 2019]. The other sectors (atmosphere, hydrosphere, biosphere) each have several RIs but EPOS has managed to bring together almost all the earth science community. With its many years of experience of FAIRness in both governance and technology, EPOS has made important contributions to ENVRI not least in leading task forces concerned with cataloguing and AAAI, both of which contribute to FAIRness, and encouraging other RIs to adopt an EPOS-like approach.

EOSC (European Open Science Cloud)[30] is a concept gradually materialising through successive projects. The concept is based on centrally managed massive distributed computing power and data storage linked to the ESFRI clusters and others as asset providers and users of assets across scientific domains. The asset base is intended to be attractive to both commercial organizations and government policymakers. EPOS has been involved in experiments with organizations of the proto-EOSC and subsequently in EOSC projects [Trani et al., 2019]. Most recently EPOS represents also ENVRI in the architecture workpackage of the latest project: EOSC-Future[31]. EPOS has mapped its metadata catalog structure to successive evolved versions of the EOSC catalog, suggesting improvements at each iteration. The EOSC architecture, built on the FAIR concept, is at the same time designed to support FAIR principles[32].

GEO is the intergovernmental partnership on earth observations[33], an initiative going beyond the pan-European landscape. Participating institutions in EPOS are active in GEO. GEOSS (GEO system of systems) [Nativi et al., 2021] is an IT solution to interoperability and there have been discussions with the EPOS IT team on the architecture and also in interoperation, as mentioned in a previous section. GEOSS is evolving to FAIRness[34] .

EGDI is the European Geological Data Infrastructure [Tulstrup et al., 2016] and is based on assets from the national geological survey organisations of Europe, many of which are partners in EPOS (e.g., BRGM – Bureau de Recherches Géologiques et Minières). A portal[35], driven by a metadata catalogue allows a user to search and find assets but not to go further such as saving into workspaces or composing workflows. The geological survey organisations are especially concerned to keep EPOS and EGDI as aligned as possible and – with some staff working on both systems – this is being achieved.

7. Future directions

Introducing FAIR principles for data provision in science has helped raising awareness on the growing problem of, on one hand, the need for interoperability of multidisciplinary data for cross-disciplinary science to solve global challenges, and on the other, the need for reusability of the wealth of scientific data for the benefit of the society. While the first two letters of FAIR, findability and accessibility, are more and more adopted by the scientific communities and the data providers, with an increased focus on open science principles, the importance of the remaining two, interoperability and reusability, are now only slowly penetrating into the scientific circles. Those are usually dominated by the ever-increasing level of specialization in science through individual PhD work focusing on minute details of a scientific problem, which are somehow distanced from the systemic approach needed to tackle

29 https://www.fairsfair.eu/ (accessed on the 12th October 2021)

30 https://eosc-portal.eu/ (accessed on the 12th October 2021)

31 https://eoscfuture.eu/ (accessed on the 12th October 2021)

32 https://www.eoscsecretariat.eu/working-groups/fair-working-group (accessed on the 12th October 2021)

33 https://www.earthobservations.org/index.php (accessed on the 12th October 2021)

34 https://www.slideshare.net/BlueBridgeVREs/how-fair-is-geoss (accessed on the 12th October 2021)

35 http://www.europe-geology.eu/onshore-geology/geological-map/ (accessed on 21st of October 2021)

the grand challenges that humanity is facing. This focus on extremely specialized knowledge is a paradox and inevitably undermines the importance of sharing data and combining data from different disciplines for solving higher level societal problems. However, in recent years it has become clear, thanks to global issues such as climate change, that the need for synthesis and systemic approach to science is now essential. In order to achieve this goal, as it is also clearly spelt out by the United Nations (UN) Sustainable Development Goals (SDG)[36], interoperability and reusability of scientific data must be granted, in addition to findability and accessibility for open science.

With the advance of data-driven science and the complexity of methods and procedures adopted in processing, analysis and simulation, reproducing scientific results becomes increasingly difficult, although this remains one of the underlying assumptions of any scientific work. FAIR should be addressed through the entire journey of data from its production to the publication through the entire data life-cycle. It is therefore desirable, in the future, further elaboration on FAIR principles so to define a comprehensive FAIR data management framework where concepts like Reproducibility of complex data products, FAIR software and data provenance information are covered. Some advances were done in this field [Hasselbring et al., 2020; Mondelli et al., 2019; Katz et al., 2021]. In particular, in [Lamprecht et al., 2019] it is remarked that many of the FAIR principles can be directly applied to research software, where software and data can be treated as the same kind of digital research objects; however, software presents specific characteristics such as its executability, composite nature, and continuous evolution accompanied by frequent versioning that make it necessary to revise and extend the original principles. Additional work is being done on defining FAIR for Research Software, in the RDA/FORCE 11/ReSA Working Group FAIR for Research Software [Gruenpeter et al., 2021], but more work is required in this direction.

An equitable focus on all aspects of FAIR principles and an improved consideration of their extension (reproducibility and FAIR software), would also pave the way for what in the experience of EPOS is becoming an urgent issue to tackle: the lack of reference architectures for implementing FAIR data stewardship systems. Although the freedom of implementation of FAIR principles is, of course, a good approach for respecting the technical history, progress and best practices of domain-specific communities, on the other hand, many data providers – especially those in need to set up a data provision system from scratch – seek for clear guidance in terms of architectures and methodologies for actual implementation. Currently, this effort is delegated to local communities and to EU-funded initiatives (e.g., ENVRI-FAIR in the environmental domain). A more systematic approach, as done for instance by the AARC initiative for the AAAI challenge at European Level (they released a blueprint architecture [AARC Community members et al., 2019]), would ease and improve the implementation of FAIR data systems. Some advances have been done in this direction by proposing the concept of FAIR implementation considerations [Katz, 2021] and a FAIR adoption process method [Bailo, 2020], however, more work is needed for the definition of a FAIR reference architectures.

The efforts made in EPOS for setting up a governance that ensures the sustainability of the entire framework, including all key players (TCS, NRI, Regional initiatives etc.) and other governance boards (e.g, IT-Board), emphasizes the importance of the definition of clear guidelines or practices followed by an organization's staff (e.g., asset access), based on policies (e.g., security, AAAI, licensing) and available through an implementation (technology). The codification of policies and making them available, the agreement on guidelines or practices for staff to follow and the correct implementation of the guidelines – respecting the policies – in IT systems is quite complex. However, appropriate policies and guidelines can improve FAIRness.

Therefore, a FAIR data management framework should include FAIR policies not only related to data assets (such as curation, provenance, security, privacy, AAAI) but also to other assets such as software (as executables or as program code for re-use), (web)services (essentially documented as an API), data products (where the derivation is recorded as provenance including the software used and configuration parameters). In these cases, much of the metadata recorded is similar to that for datasets but there are some differences. Also licensing conditions for software may be different from those for data to ensure continued availability and openness.

36 https://www.africanpromise.org.uk/charity-work/supporting-the-sustainable-development-goals/?gclid=Cj0KCQjwtrSLBhCLARIsACh6RmjHa9CfOirBYFhBJrpqQ60uRlhy9gJ9qmZfFLB0LYzClGq6QnZspokaAtADEALw_wcB (accessed on the 12th October 2021)

8. Conclusions

In the current work, we described the EPOS approach to Data integration and FAIR data management. The complexity of the EPOS framework, which includes Thematic Communities as data providers (TCS), a Central Hub integration node (ICS), distributed nodes (ICS-D) for data processing or visualization and the users, is exposed in detail emphasizing the heterogeneity of assets in terms of data, data products, software and services (DDSS) from more than 250 data providers, grouped in 10 sub-disciplines, in the domain of Solid Earth Sciences. To reflect the complexity of the data organization, a data taxonomy was conceived and data were also grouped into three main different categories: georeferenced data (e.g., maps), time series, that describe the temporal evolution of a specific physical dimension, and non geo-referenced data, as software packages or list of pdf reports.

To integrate such heterogeneity, EPOS was considered as a prominent example of Information-Powered Collaborations, where FAIR principles can support the shaping of the technical underpinning e-infrastructure by targeting key requirements of findability, accessibility, interoperability and reusability. However, in order to keep consistency with the other dimension of the EPOS framework (e.g., sustainability and governance), the IT design had to keep a constant communication with the other non-technical aspects to ensure that the technical architecture was coherent with governance, financial aspects, legal aspects and communities' organizations. This is referred to as FAIR governance framework.

For the technical integration of the assets provided by the communities, EPOS relied upon two main elements that were recognized as necessary: the metadata and the service-based software architecture. The best approach to interoperation was recognized to be a hybrid approach that took the best out of three methods that were investigated: brokering, metadata and web services approach. The hybrid approach means that brokers convert from many metadata formats, exposed to the consumers as web services, to one central system (ICS-C) that uses a rich metadata catalogue.

By adopting such an approach, a multi-tier application for accessing multi-disciplinary data from solid Earth science domain was built: the EPOS Data portal[37]. It includes a Graphic User Interface by means of which users can browse services, dig down to selected dataset slices, pre-visualize data in overlap mode and in three different modes (map, time-series, tabular), and eventually collect selected data in a personal workspace. A prototype of distributed service (ICS-D) for advanced visualization and processing has also been developed.

EPOS approach to integrated data management has been developed since EPOS' early stages (EPOS-IP in 2011), and has – to some extent – foreseen the technical implications of FAIR. The emphasis on rich metadata catalogue and service-based software architecture, combined with a FAIR adoption roadmap relying upon four conceptual layers guiding the architecture design and implementation (data, metadata, access services, usage services), together with best practices on crucial aspects (metadata, semantics, interoperability, AAI, community building), have enabled EPOS to be compliant with the FAIR principles.

The collaborative endeavour for building an integrated system for FAIR data access, carried out by a huge multi-disciplinary community of practitioners, engineers and scientists for almost a decade, also in synergy with other initiatives (e.g., ENVRI-FAIR, GEO, GEOSS, EOSC etc.), allowed us to identify key challenges for the FAIR community.

Up to now the focus has been put mostly on the first two letters of FAIR for enabling true open science, while the remaining two are taking time to penetrate into the scientific daily practices. However, the need for synthesis and systemic approach to science makes the adoption of appropriate interoperability and reusability practices more and more urgent. Also, a FAIR management framework addressing FAIR through the entire data lifecycle, fostering best practices for Reproducibility of scientific results, FAIR software and data provenance, is envisaged as an open path to true FAIRness.

In frameworks like EPOS, where individuals with different background, goals and roles are required to collaborate in the construction of data stewardship system, the freedom of implementation of FAIR principles is an efficient approach because it respects the technical history, progress and best practices of domain-specific communities. However, especially for those who need to set up a data provision system from scratch, there is a clear need for guidance in the choice of architectures and methodologies, now delegated to local communities and EU-funded initiatives (e.g., ENVRI-FAIR) with the risk of leading to incompatible implementation of FAIR data stewardship

37 https://www.ics-c.epos-eu.org/ (accessed on the 21st of October 2021)

systems. Different approaches, providing pragmatic FAIR implementation consideration [Jacobsen et al., 2020], catalogue of existing implementation solutions as in the case of the FAIR convergence Matrix [Sustkova et al., 2020], FAIR adoption methods and – most importantly – FAIR reference architectures, are envisaged to facilitate and speed up the construction of FAIR data systems.

Finally, the EPOS enterprise has also shown that FAIR data can be provided in a sustainable way only when FAIR governance and FAIR data policies frameworks are adopted at a wide level in the community, in addition to the mere technical work.

The above emphasizes the importance of Research Infrastructures like EPOS, as the place where reusable open science practices are defined, implemented, and can be shared, and as the most appropriate frameworks where sustainability can be pursued for maintaining and operating FAIR services. Research Infrastructures are the place where a strong relationship with the scientists and data providers is ensured, and where governance, policies, best practices and reference architectures are adopted in a harmonized way within a community, thus making FAIR principles a reality and improving the Open Science in the day-to-day work of scientists in a sustainable way. The important role of Research Infrastructures in the FAIR community, that are building operational open-science oriented systems, should be therefore strongly emphasized and taken into account at all levels of the wider EU planning action.

Acknowledgements. The authors would like to acknowledge the following projects and initiatives for financing part of the activities: EPOS-IP, European Commission H2020 program, Grant agreement ID 676564; EPOS-ERIC (https://www.epos-eu.org/epos-eric) for sustainability and management; NORCE for ICS-D prototype development.
The Authors would also like to acknowledge the developers from the EPOS IT Team for taking part in development activities related to the architecture, GUI implementation, ICS-D prototype development. In particular: the EPOS BGS IT Team for development activities related to the architecture and GUI implementation and design, and specifically Patrick D. Bell, Christopher D. Card, Jon Stuteley, Phil Atkinson, Daniel L.W. Warren; the EPOS INGV IT team, for development activities related to the architecture, GUI implementation and design and metadata, especially Lorenzo Fenoglio, Sara Capotosti and Manuela Sbarra; NORCE team for the development of the ICS-D prototype "Enlighten", in particular Tor Langeland and Ove Lampe; the EPOS-ECO Communication unit, in particular Barbara Angioni.

References

AARC Community members, & AppInt members (2019). AARC Blueprint Architecture 2019 (AARC-G045). Zenodo. https://doi.org/10.5281/zenodo.3672785

Atkinson, M., R. Filgueira, I. Klampanos, A. Koukourikos, A. Krause, F. Magnoni, C. Pagé, A. Rietbrock, A. Spinuso (2019). Comprehensible control for researchers and developers facing data challenges. In *2019 15th International Conference on eScience (eScience)*, 311-320, IEEE, https://doi.org/10.1109/eScience.2019.00042

Bailo, D., R. Paciello, M. Sbarra, R. Rabissoni, V. Vinciarelli, M. Cocco (2020). Perspectives on the Implementation of FAIR Principles in Solid Earth Research Infrastructures, Front. Earth Sci., 8, 3, https://doi.org/10.3389/feart.2020.00003

Bailo, D., and K. G. Jeffery, (2014). EPOS: a novel use of CERIF for data-intensive science, Proc. Comput. Sci., 33, 3-10, https://doi.org/10.1016/j.procs.2014.06.002.

Bailo, D., K. G. Jeffery, A. Spinuso, G. Fiameni (2015). Interoperability oriented architecture: the approach of EPOS for solid Earth e-infrastructures, In 2015 IEEE 11th International Conference on e-Science, 529-534, https://doi.org/10.1109/eScience.2015.22

Candela, L., D. Castelli and P. Pagano (2013). Virtual Research Environments: An Overview and a Research Agenda, Data Sci. J., 12, 0), GRDI75-GRDI81, https://doi.org/10.2481/dsj.GRDI-013

Chiaraluce, L. (2012). Unravelling the complexity of Apenninic extensional fault systems: A review of the 2009 L' Aquila earthquake (Central Apennines, Italy), J. Struct. Geol., 42, 2-18. https://doi.org/10.1016/j.jsg.2012.06.007

Collins, S., F. Genova, N. Harrower, S. Hodson, S. Jones, L. Laaksonen, D. Mietchen, R. Petrauskaité and P.Wittenburg (2018). Turning FAIR data into reality: interim report from the European Commission Expert Group on FAIR data. Interim Report from the European Commission Expert Group on FAIR Data, June. https://doi.org/10.5281/zenodo.1285272

Collins, S., F. Genova, N. Harrower, S. Hodson, S. Jones, L. Laaksonen and P. Wittenburg (2018). Turning FAIR into reality: Final report and action plan from the European Commission expert group on FAIR data.

Giardini, D., J. Wössner and L. Danciu (2014). Mapping Europe's Seismic Hazard. Eos, Trans. Am. Geophys. U., 95, 29, 261-262, https://doi.org/10.1002/2014EO290001

Gruenpeter, M., D.S. Katz, A-L. Lamprecht, T. Honeyman, D. Garijo, A. Struck, A. Niehues, P. A. Martinez, L. J. Castro, T. Rabemanantsoa, N. C. Hong, P. Neil, C. Martinez-Ortiz, , L. Sesink, M. Liffers, A. C. Fouilloux, C. Erdmann, S. Peroni, P. Martinez Lavanchy, , I. Todorov, M. Sinha (2021). Defining Research Software: a controversial discussion (Version 1). Zenodo, https://doi.org/10.5281/zenodo.5504016

Hasselbring, W., L., Carr, S. Hettrick, H. Packer, T. Tiropanis (2020). From FAIR research data toward FAIR and open research software, IT - Info. Tech., 62, 1, 39-47, https://doi.org/10.1515/itit-2019-0040.

Hong, N.C., S. Cozzino, F. Genova, M. Hoffman-Sommer, R. Hooft, L. Lembinen, J. Marttila, M. Teperek (2020). Six Recommendations for implementation of FAIR practice by the FAIR in practice task force of the European open science cloud FAIR working group (Issue October), https://doi.org/10.2777/986252

Jacobsen, A., R. de Miranda Azevedo, N. Juty, D. Batista, S. Coles, R. Cornet, M. Courtot, M. Crosas, M. Dumontier, C. T. Evelo, C. Goble, G. Guizzardi, K. K. Hansen, A. Hasnain, K. Hettne, J. Heringa, R. W. W. Hooft, M. Imming, K. G. Jeffery, Schulte (2020). FAIR Principles: Interpretations and Implementation Considerations, Data Intelligence, 2, 1-2, 10-29, https://doi.org/10.1162/dint_r_00024

Jeffery, K., N. Houssos, B. Jörg, A. Asserson (2014): Research Information Management: The CERIF Approach, Int. J. Metadata, Semantics and Ontologies, 9, 1, 5-14, https://doi.org/10.1504/IJMSO.2014.059142

Jeffery, K. G. and D. Bailo (2014). EPOS: using metadata in geoscience. In Research Conference on Metadata and Semantics Research, 170-184, Springer, Cham., https://doi.org/10.1007/978-3-319-13674-5_17

Katz, D.S., M. Barker, N. P. C. Hong, L. J. Castro and P.A Martinez (2021). The FAIR4RS team: Working together to make research software FAIR. 2021 Collegeville Workshop on Scientific Software – Software Teams (Collegeville2021), https://doi.org/10.5281/zenodo.5037157

Klampanos, I., A. Davvetas, A. Gemünd, M. Atkinson, A. Koukourikos, R. Filgueira, A. Krause, A. Spinuso, A. Charalambidis, F. Magnoni, E. Casarotti, C. Pagé, M. Lindner, A. Ikonomopoulos and V. Karkaletsis (2019). DARE: A reflective platform designed to enable agile data-driven research on the cloud. In 2019 15th International Conference on eScience (eScience), 578-585, https://doi.org/10.1109/eScience.2019.00079

Koers, H., D. Bangert, E. Hermans, R. van Horik, M. de Jong and M. Mokrane (2020). Recommendations for Services in a FAIR Data Ecosystem, Patterns, 1, 5, https://doi.org/10.1016/j.patter.2020.100058.

Lamprecht, A.-L., L. Garcia, M. Kuzak,., C. Martinez, R. Arcila, E. Martin Del Pico, V. Dominguez Del Angel, S. van de Sand, J. Ison, P. A. Martinez, P. McQuilton, A. Valencia, J. Harrow, F. Psomopoulos, J.L. Gelpi, N. C. Hong, C. Goble and S. Capella-Gutierrez (2019). Towards FAIR principles for research software. Data Science, 3(1), 37-59, https://doi.org/10.3233/ds-190026.

Langeland, T., O. D., Lampe, G. Fonnes, K. Atakan, J. Michalek, X. Wang, C. Rønnevik, T. Utheim, K. Tellefsen (2019). EPOS-Norway Portal. In Geophysical Research Abstracts, 21.

Michener, W. K., S. Allard, A. Budden, R. B. Cook, K. Douglass, M. Frame, S. Kelling, R. Koskela, C. Tenopir and D. A. Vieglais, (2012). Participatory design of DataONE-Enabling cyberinfrastructure for the biological and environmental sciences, In Ecological Informatics, 11, 5-15, https://doi.org/10.1016/j.ecoinf.2011.08.007.

Mondelli, M. L., A. Townsend Peterson and L. M. R Gadelha (2019). Exploring reproducibility and FAIR principles in data science using ecological niche modeling as a case study. Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 11787 LNCS, Ml, 23-33, https://doi.org/10.1007/978-3-030-34146-6_3.

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

Nativi, S., P. Mazzetti, M. Craglia and N. Pirrone (2014). The GEOSS solution for enabling data interoperability and integrative research, Environ. Sci. Pollut. Res., 21, 4177-4192, https://doi.org/10.1007/s11356-013-2264-y.

Nativi, S., K.G. Jeffery, R. Koskela (2015). RDA: Brokering with Metadata; ERCIM News 100 January 2015, 25-26.

Nativi, S. and M. Cragli (2021). Evolution of GEOSS Common Infrastructure, European Commission, Ispra, 2021, JRC123141.

Newman, Building Microservices (2015). O'Reilly Media. ISBN: 9781491950357.

Orlecka-Sikora, B., S. Lasock, J. Kocot et al. (2020). An open data infrastructure for the study of anthropogenic hazards linked to georesource exploitation, Sci. Data, 7, 89, https://doi.org/10.1038/s41597-020-0429-3.

Paciello, R., L. Trani, D. Bailo, V. Vinciarelli, M. Sbarra (2021). SHAPEness: a SHACL-driven RDF Graph Editor, Semantic Web – Interoperability, Usability, Applicability an IOS Press Journal, (under review), http://www.semantic-web-journal.net/content/shapeness-shacl-driven-rdf-graph-editor-0.

Petzold A., A, Asmi, A. Vermeulen, G. Pappalardo, D. Bailo, D. Schaap et al. (2019), ENVRI-FAIR – Interoperable Environmental FAIR Data and Services for Society, Innovation and Research, in 2019 15th International Conference on eScience (eScience), 277-280, https://doi.org/10.1109/eScience.2019.00038.

Richardson, C. (2017). Microservices Patterns. In Online, 129, http://www.ncbi.nlm.nih.gov/pubmed/20608803

Richardson, L., and S. Ruby (2007). RESTful Webservices. O'Reilly Media, Inc.

Sansone, S. A., P. McQuilton, P. Rocca-Serra, A. Gonzalez-Beltran, M. Izzo, A. L. Lister and M. Thurston (2019). FAIRsharing as a community approach to standards, repositories and policies, Nature biotechnol., 37, 4, 358-367, https://doi.org/10.1038/s41587-019-0080-8

Schuldt, H. (2009). Multi-Tier Architecture in Encyclopedia of Database Systems, https://doi.org/10.1007/978-0-387-39940-9_652

Schultes, E. A., G. O Strawn and B. Mons (2018). Ready, Set, GO FAIR: Accelerating Convergence to an Internet of FAIR Data and Services in Proceedings of the XX International Conference "Data Analytics and Management in Data Intensive Domains" (DAMDID/RCDL'2018), 19-23, Moscow, Russia, October 9-12, 2018.

Sinaeepourfard, A., X. Masip-Bruin, J. Garcia and E. Marín-Tordera (2015). A Survey on Data Lifecycle Models: Discussions toward the 6Vs Challenges, Technical Report (UPC-DAC-RR-2015-18).

Spinuso, A, M. Veldhuizen, D. Bailo, V. Vinciarelli, T. Langeland (2022). SWIRRL Managing Provenance-Aware and Reproducible Workspaces, Data Intelligence Journal on Canonical Workflow Frameworks for Research, in press.

Spinuso, A., I. van der Neut, H. Verhoef, S. Wagenaar and F. Striewski (2020). SWIRRL: Software for Interactive and Reproducible Research Labs, in IWSG 2020 Workshop, Zenodo. https://doi.org/10.5281/zenodo.4264852

Stall, S., L. Yarmey, J. Cutcher-Gershenfeld, B. Hanson, K. Lehnert, B. Nosek, M. Parsons, E. Robinson and L. Wyborn (2019). Make scientific data FAIR, Nature, 27-29, https://doi.org/10.1038/d41586-019-01720-7.

Sustkova, H. P., K. M. Hettne, P. Wittenburg, A. Jacobsen, T. Kuhn, R. Pergl, J. Slifka., P. McQuilton, B. Magagna, S. A. Sansone, M. Stocker, M. Imming, L. Lannom, M. Musen and E. Schultes (2020). Fair convergence matrix: Optimizing the reuse of existing fair-related resources, Data Intelligence, 2, 1-2, 158-170, https://doi.org/10.1162/dint_a_00038.

Trani, L. (2019). A Methodology to Sustain Common Information Spaces for Research Collaborations, University of Edinburgh.

Trani, L., M. Fares, J. Paulo, P. Zanetti, J. Quinteros, N. Triantafyllis (2019). Building the EPOS-ORFEUS Competence Center in EOSC-hub, https://doi.org/10.13140/RG.2.2.33146.75205.

Trani, L., M. Atkinson, D. Bailo, R. Paciello, R. Filgueira (2018). Establishing Core Concepts for Information-Powered Collaborations, Futur. Gener. Comput. Syst., 89, 421-437. https://doi.org/10.1016/j.future.2018.07.005

Tulstrup, J., A. Tellez-Arenas, M. Pedersen, F. Robida, B. Pjetursson and C. Delfini (2016). The European Geological Data Infrastructure EGDI, in 35th International Geological Congress: IGC 2016.

Vicente-Saez, R., C. Martinez-Fuentes (2018). Open Science now: A systematic literature review for an integrated definition, J. Bus. Res., 428-436, https://doi.org/10.1016/j.jbusres.2017.12.043.

Wilkinson, M. D., M. Dumontier, J. Ij.Aalbersberg, G. Appleton, M. Axton, A. Baak, et al. (2016). The FAIR guiding principles for scientific data management and stewardship, Sci. Data, 3, 160018, https://doi.org/10.1038/sdata.2016.18.

***CORRESPONDING AUTHOR: Daniele BAILO**,
Istituto Nazionale di Geofisica e Vulcanologia (INGV),
email: daniele.bailo@ingv.it