Data Management in Distributed, Federated Research Infrastructures: The Case of EPOS

PRACTICE PAPER

DANIELE BAILO
ROSSANA PACIELLO
JAN MICHALEK
DANIELA MERCURIO
AGATA SANGIANANTONI
KAUZAR SALEH CONTELL
OTTO LANGE
GIOVANNA MARACCHIA
KUVVET ATAKAN
KEITH G. JEFFERY
CARMELA FREDA

*Author affiliations can be found in the back matter of this article

ABSTRACT

Data management is a key activity when Open Data stewardship through services complying with the FAIR principles is required, as it happens in many National and European initiatives. Existing guidelines and tools facilitate the drafting of Data Management Plans by focusing on a set of common parameters or questions. In this paper we describe how data management is carried out in EPOS, the European Research Infrastructure for providing access to integrated data and services in the solid Earth domain. EPOS relies on a federated model and is committed to remain operational in the long term. In EPOS, five key dimensions were identified for the Federated Data Management, namely the management of: thematic data; e-infrastructure for data integration; community of data providers committed to data provision processes; sustainability; and policies. On the basis of the EPOS experience, which is to some extent applicable to other research infrastructures, we propose additional components that may extend the EU Horizon 2020 Data Management Guidelines template, thus comprehensively addressing the Federated Data Management in the context of distributed Research Infrastructures.

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

KEYWORDS:
Data Management; Federated Research Infrastructures; EPOS; FAIR Principles; Open Data; Data Integration; Solid Earth Science; Data Management Plans

TO CITE THIS ARTICLE:
Bailo, D, Paciello, R, Michalek, J, Mercurio, D, Sangianantoni, A, Contell, KS, Lange, O, Maracchia, G, Atakan, K, Jeffery, KG and Freda, C. 2024. Data Management in Distributed, Federated Research Infrastructures: The Case of EPOS. *Data Science Journal*, 23: 5, pp. 1–17. DOI: https://doi.org/10.5334/dsj-2024-005

1. INTRODUCTION

For almost a decade, the EU demands the use of Data Management Plans (DMP) for describing the strategies applied when managing data throughout their lifecycle (European Commission 2013). Since DMPs generally describe the solutions chosen to meet optimized levels of sustainable and FAIR data management, they must be considered as key elements for understanding how an initiative is supposed to foster Open Science and to allow the re-use of digital assets. Compared to the past, these DMPs represent a significant step forward as they provide an easy way to clarify how data will be stored, maintained, curated, etc. Although they generally suit very well to time-limited initiatives, like EU projects, currently available templates for DMPs present interesting challenges when applied to different contexts, for instance when the data from heterogeneous data sources is provided in an integrated way in a federated environment for fostering FAIR data access through an Open Data system. Such a case is indeed not uncommon in many Research Infrastructures, as those included in the European Strategic Forum on Research Infrastructures (ESFRI) roadmap (ESFRI 2018), like ICOS[1] (Vermeulen et al. 2015) in the domain of greenhouse gas measurements, EMSO[2] (Best et al. 2016) in the Oceanography domain, EPOS[3] (Cocco et al. 2022a) in the domain of solid Earth, and many others which are based upon a federated model. In this case, addressing the technical dimension of data management is not sufficient, as it is inevitably intertwined with the governance underlying the integrated data provision, the sustainability dimension and – when the Research Infrastructure is required to be operational under clear commitments – also with the financial and legal dimensions. Aspects going beyond the technical dimension are considered by existing guidelines, as for instance the EU Horizon 2020 Guidelines (European Commission 2013), but still a gap exists with respect to current practices in Federated Research Infrastructures. In the case of EPOS, a distributed Research Infrastructure relying on a federated data provision approach, the management of Data, Data products, Software and Services required tackling several different dimensions and to take account of the distributed nature of the entire network of data providers. Such experience is somehow iconic and representative of other RIs, for instance those in the Environmental domain, and can potentially contribute to the Horizon 2020 Data Management Plan guidelines (European Commission 2016). In the remainder of the paper, we describe the approach of EPOS to such a Federated Data Management and we provide a first set of suggestions for improving the existing Horizon 2020 Data Management Plan guidelines template.

2. BACKGROUND

DATA MANAGEMENT PLANNING

Data Management Plans were introduced in the Horizon 2020 Work Programme for 2014–2015 with the purpose of 'detailing what data the project will generate, whether and how it will be exploited or made accessible for verification and re-use, and how it will be curated and preserved' (European Commission 2013). Such a request was motivated by the recognition that research data is 'as important as the publications it supports, and by the concerns related to the availability of data produced in the framework of a project, even after the project ended.'

The core elements of the Data Management Plan (DMP) include dataset name and description, metadata standards used for describing the dataset, data stewardship mechanisms (including technical details and used standards), Archiving and long-term preservation procedures, as highlighted in the Guidelines on Data Management in Horizon 2020 document.

Since then, a number of documents have been developed for supporting researchers in drafting DMPs, as in the cases of the guidelines released by the University of Bristol, focused

[1] ICOS improves the quality, spatial resolution, and time-series length of greenhouse gas observations by providing provide an added value of data from three domains (atmosphere, ecosystem, ocean) from a network of 13 countries with more than 150 stations https://www.icos-cp.eu/ (accessed on 10 December 2022).

[2] The European Multidisciplinary Seafloor and water column Observatory (EMSO) aims to explore the oceans, to gain a better understanding of phenomena happening within and below them, and to explain the critical role that these phenomena play in the broader Earth system. https://emso.eu/ (accessed on 10 December 2022).

[3] EPOS, the European Plate Observing System, is a multidisciplinary, distributed research infrastructure that facilitates the integrated use of data, data products, and facilities from the solid Earth science community in Europe. https://www.epos-eu.org/ (accessed on 10 December 2022).

on supporting researchers in drafting DMP for research projects (Bristol University 2014), and of those from the University of New Mexico (US) where DMPs are considered in relation to the Data Lifecycle for a simple use case (Michener 2015), and the key 43 topics to be dealt with in a DMP are identified (Williams, Bagwell & Nahm Zozus 2017). Likewise, interesting tools have been developed for facilitating the production of DMPs, as in the case of DMPTool, a 'free, open source, online tool that helps users build a comprehensive and descriptive data management plan' (Swauger 2015) and of DMP Online, a tool to 'assists researchers, data custodians and other stakeholders in creating, maintaining and exporting data management plans' (Donnelly, Jones & Pattenden-Fail 2010; Sallans & Donnelly 2012).

More recently, DMP requirements, grouped in 14 main themes, have been related to existing solutions (Jones et al. 2020), contextualized in an ecosystem of tools for fostering the implementation of FAIR principles (Wilkinson et al. 2016). Interestingly, this publication also leveraged the work done in the framework of Research Data Alliance – RDA.[4] Such efforts represent an attempt to bridge the gap between the planning, the assessment, the compliance to FAIR principles – which are now recognized by the EU as common core criteria – and the latest evolution of DMPs as machine-actionable DMPs (Miksa et al. 2018).

The literature on DMP is of course huge and it is out of the scope of this work to provide an overall and comprehensive view of the work done so far on the topic. However, as of the day of writing, apart some contributions in the form of reports (Karasti et al. 2018), burning questions about how to implement Data Management Planning in the context of huge, Federated Research Infrastructures – as those included in the ESFRI Roadmap (ESFRI–European Strategy Forum on research Infrastructures 2021) – presenting a high heterogeneity in terms of data formats, standards, stewardship systems and policies – still remain an open ground for research. In this framework, the next section introduces the case of the EPOS solid Earth Sciences Research Infrastructure, presenting an approach of Data Management Planning in a federated, heterogeneous context.

THE EPOS FEDERATED APPROACH

The European Plate Observing System (EPOS),[5] is a multidisciplinary, distributed research infrastructure that facilitates the multidisciplinary use of data, data products and facilities from the solid Earth science community in Europe (Cocco et al. 2022b) by integrating diverse European Research Infrastructures under a common federated framework. EPOS is the sole research infrastructure representing solid Earth science in Europe.

EPOS initiatives exist for more than a decade, firstly as a project included into the European Strategic Forum on Research Infrastructures (ESFRI) Roadmap in 2008, then implemented through two European Projects, namely EPOS-PP Preparatory Phase (2010–2014), Grant Agreement no. 262229, and EPOS-IP Implementation Phase (2015–2019), Grant Agreement no. 676564, and finally established as an European Research Infrastructure Consortium (ERIC) in 2018, a specific legal form that facilitates the establishment and the operation of RIs with pan-European dimensions.

The EPOS Data Portal[6] provides homogenous access to heterogeneous resources made available by data providers in the EPOS community. EPOS adopted a specific methodology to comply with the FAIR principles which relies on a federated approach for providing access to Data, Data Products Services and Software (Bailo, Jeffery et al. 2022), and also faced ethical challenges related to data collection, data use as well as with respect to their promotion (Marti et al. 2022).

The main technical challenge in the EPOS case was to integrate more than 250 services, delivering more than 30 different types of data formats – spanning for instance from miniSeed waveforms for Seismology to tiff images for Satellite Data and pdf reports for Volcanology – covering more than 800 TB of data in total, described by more than 20 different types of metadata. Some metadata is following international or community-based standards, but some had to be designed and developed from scratch. This process was speeded up thanks to sharing of best practices among individual communities in EPOS.

[4] RDA-DMP-Common-Standard, https://github.com/RDA-DMP-Common/RDA-DMP-Common-Standard (accessed on 14.12.2022).

[5] https://www.epos-eu.org/ (accessed on 10 December 2022).

[6] https://www.epos-eu.org/dataportal (accessed on 10 December 2022).

The EPOS architecture is based on a three-component structure (Figure 1). The first component consists of the main data providers that collect solid Earth data through various National Research Infrastructures (NRIs), that maintain their integrity with respect to data ownership, storage and availability. Data collected by the NRIs are then combined at the middle layer (Thematic Core Services – TCS), where community-based standards are developed and data are made available to the entire European scientific community (e.g., ORFEUS[7] in seismology). Services provided through these TCS form the basis for the third component where resources are integrated and made interoperable in a metadata structure. This component represents the EPOS Integrated Core services (ICS) with the associated EPOS Data Portal. The EPOS Delivery Framework refers to the TCS, ICS and ECO, as these are the elements included in the EPOS governance model.

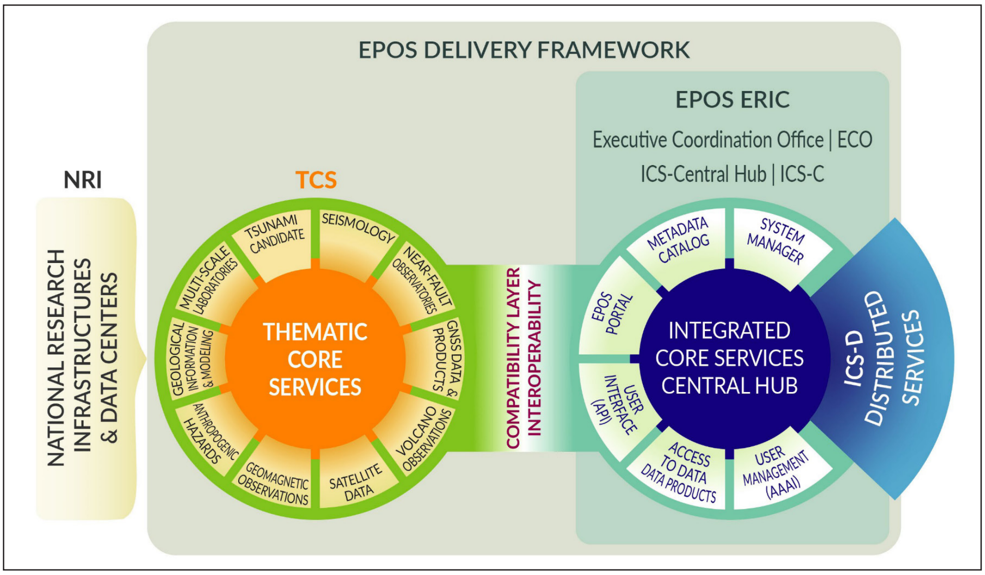


Figure 1 Main elements of the EPOS Architecture: National Research Infrastructures (NRI), Thematic Core Services (TCS), Integrated Core Services Centralized and Distributed (ICS-C & ICS-D) form the EPOS functional architecture, coordinated by EPOS ERIC, and designed to ensure the EPOS data and service provision. The Integrated Core Services Central Hub (ICS-C) and the Executive and Coordination Office (ECO) belong to the EPOS ERIC legal subject.

The EPOS architecture therefore encompasses several research infrastructures at different levels (local, regional, national, transnational, and pan-European) with the goal of harmonizing and integrating the data provision; such complexity needs to be reflected also in the Data Management Plans and strategies.

Strategies and practices for addressing community specific DMPs, as well as the DMP for the Integration Infrastructure are outlined in the next section, where also a multidimensional approach for Federated Data Management in EPOS is described.

3. SOLID EARTH SCIENCE FEDERATED DATA MANAGEMENT IN EPOS

In the EPOS federated approach, the DMP harmonization process is not implemented as a simple mapping of Data Management procedures and aspects that are common across all communities, or as a huge volume including DMPs from all communities, which would not reflect the harmonization efforts. Each of the thematic communities indeed offers access to data, data products and services that vary significantly and are provided by a large number of institutions that constitute the underlying NRIs, each one having its own DMP. In addition, not all TCS have developed DMPs yet, as some of these are being incrementally developed and there is indeed a need for harmonizing them at TCS level.

The approach taken in the EPOS community needs therefore to take into account two key aspects: the first one is the just discussed heterogeneity, the second one is related to the status of the operational Research Infrastructure.

As for the heterogeneity, this can be addressed by a) ensuring each NRI or TCS has its own DMP, b) by releasing a DMP for the EPOS Data Portal and the underpinning e-Infrastructure (ICS-C) and finally c) by providing EPOS DMP Guidelines so to harmonize DMPs provided by the Thematic Communities, as already done in the EPOS-IP phase (Atakan et al. 2019, EPOS IP WP6 and WP7 Teams 2015).

[7] http://www.orfeus-eu.org/ (accessed on 10 December 2022).

As for the status of Operational RI, EPOS is committed to remain an operational Research Infrastructure with a well-defined legal form, the aforementioned ERIC, where clear commitments are made with respect to the sustainability of the (integrated) data provision. This requires the adoption of a multi-dimensional approach that goes beyond the ordinary themes of the DMP as described in previously mentioned works, e.g., (Michener 2015) and (Williams, Bagwell & Nahm Zozus 2017), and that accounts for the Community, Governance, Policy and Sustainability dimensions, which are therefore effectively part of the Data Management Planning in a federated landscape like EPOS.

DATA MANAGEMENT OF THEMATIC DATA

The Thematic Core Services (TCS) integrate diverse communities. There are currently nine Thematic Communities[8] as seismology, GNSS data, volcano observations and geological modelling, with tsunami as candidate Thematic Community. They all act as transnational governance frameworks where data and services are provided to answer scientific questions, and where each community discusses a range of issues that spans from implementation to ethics. The Thematic nodes are interoperable with the Integrated Core Services (ICS) via the Interoperability Layer, thus making their datasets and services available through the EPOS Data Portal (Figure 1).

For each community, a first round of DMP collection was conducted during the EPOS-IP project (2015–2019). Community DMPs were then harmonized and guidance was provided to the communities about how to manage specific issues, including data policy, technical issues (e.g., architectures, software solutions and approaches for data preservation and storage), ethical and governmental themes, as discussed elsewhere (Atakan et al. 2019, EPOS IP WP6 and WP7 Teams 2015). To align with the requirements stemming from the FAIR principles and from the EU DMP recommendations for Horizon 2020 (European Commission 2016), the official DMP guidelines available at the time, guidelines for the communities DMP were agreed, which included six sections:

1. Data Generalities, including a brief description of the data, type of data and its coverage must be provided. Data formats, origins, total expected size of datasets and type of users should also be outlined.
2. FAIR data, which is all about making data findable (including provisions for metadata), making data openly accessible, interoperable, and increase data re-use.
3. Allocation of Resources, describing the resources to make data FAIR and to ensure its sustainability, as well as resources needed to deliver the plan, e.g., software, hardware, technical expertise, financial, funding, etc.
4. Data Security, address data recovery as well as secure storage and transfer of sensitive data.
5. Ethical Aspects, addressing ethical or legal issues that can have an impact on data sharing.
6. Other Issues, addressing any additional information about other national/funder/sectorial/departmental procedures for data management currently in place.

Taking into account the need for a governance to oversee the relationships among various data contributors – namely National Research Infrastructures, Service Providers, thematic nodes – the guidelines influenced the creation of the Thematic Core Services 'Consortium Agreements.' These agreements, finalized in 2018–2019, facilitate the establishment, administration, governance, and operation of the TCS creating the legal condition to link the TCS to EPOS ERIC. Due to the legal nature of the agreements and their legal requirements, the data management description differs from the templates provided by the EU in 2016 and also from the more recent templates included in the Horizon Europe Guidelines (European Commission 2021). Nevertheless, the agreements retain most of the key concepts from the European Commission DMPs template. For example, they integrate the EPOS Data Policy Management as an Annex, introducing the principles and process of handling data and intellectual property rights within EPOS activities.

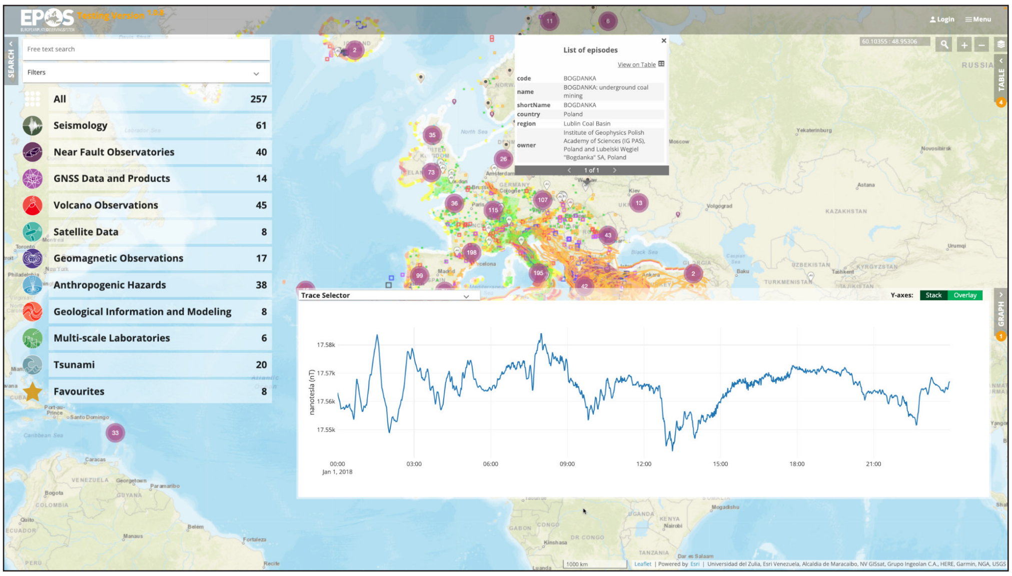
The complexity of a harmonized Data Management in a federated environment therefore goes beyond a simple 'DMP harmonization,' and clearly requires addressing legal aspects, both part of the Sustainability dimension, in a community driven fashion, guaranteeing a difficult balance

[8] TCS full list is the following: seismology, near fault observatories, GNSS data and products, volcano observations, satellite data, geomagnetic observations, anthropogenic hazards, geological information and modeling, multi-scale laboratories, tsunami, also available here: https://www.epos-eu.org/tcs (Accessed on 1 September 2023).

that ensures harmonized governance and harmonized legal agreements across communities and – at the same time – respect to the specific community needs and story.

MANAGEMENT OF THE DATA INTEGRATION INFRASTRUCTURE

Managing heterogeneous data that is distributed among diverse sources, which can be organized, semi-structured or unstructured, and giving users a uniform access to such data is fundamental to achieve a true data integration. In response to such a requirement, EPOS built the EPOS Data Portal (Figure 2), which represents a one-stop shop aiming at facilitating data access and at providing advanced features for exploration, exploitation, and re-usability of heterogeneous data across different scientific disciplines in solid Earth science. The EPOS technical architecture enables to collect, integrate, and efficiently manage information to provide homogeneous access over heterogeneous data through a four-tier architecture (Bailo, Paciello et al. 2022). This should not be confused with the EPOS functional architecture of Figure 1, with which it shows some common elements. The technical architecture includes:



All	257
Seismology	61
Near Fault Observatories	40
GNSS Data and Products	14
Volcano Observations	45
Satellite Data	8
Geomagnetic Observations	17
Anthropogenic Hazards	38
Geological Information and Modeling	8
Multi-scale Laboratories	6
Tsunami	20
Favourites	8

List of episodes

View on Table

code	BOGDANKA
name	BOGDANKA: underground coal mining
shortName	BOGDANKA
country	Poland
region	Lublin Coal Basin
owner	Institute of Geophysics Polish Academy of Sciences (IG PAS), Poland and Lubelski Węgiel "Bogdanka" SA, Poland

1 of 1

Figure 2 *The EPOS Data Portal.* Managing Data provision through the ICS-C system also required addressing the topics of policies and sustainability, detailed later in this paper.

Tier 1, Thematic Core Services (TCS), representing European wide technical services providing access to datasets and services by domain specific communities.

Tier 2, Interoperability layer, aiming at collecting descriptions of TCS resources (e.g., Dataset or Services) provided by TCS through a common knowledge representation language, namely EPOS-DCAT-AP (Paciello et al. 2022).

Tier 3, Integrated Core Services, which includes a Central Hub system (ICS-C) and Distributed systems (ICS-D), are responsible for aggregating and harmonizing TCS assets as well as integrating external computational and processing services.

Tier 4, Graphical User Interface (GUI), provides access to multidisciplinary data by offering different and multiple ways for users to discover, contextualize and download data.

The complexity inherent to the Data Portal development and maintenance – not only in technical terms, but also from a human perspective – entails challenges that are addressed by adopting a co-development approach and continuous integration techniques. The adoption of such techniques and dedicated tools (e.g., GitLab,[9] GitHub[10]) together with constant DevOps meetings, are crucial to automate complex code integration and error-checking, as well as to streamline communication between the developers themselves and related teams.

[9] https://gitlab.com (accessed on 14 December 2022).

[10] https://github.com (accessed on 14 December 2022).

Managing Data provision through the ICS-C system also required addressing the topics of policies and sustainability, detailed later in this paper.

MANAGEMENT OF THEMATIC COMMUNITIES FOR INTEGRATED DATA PROVISION

One of the relevant aspects in managing the Data Integration Infrastructure is the co-development approach. IT enabled data providers, developers, and operator teams to continuously guide the evolution of the EPOS Data Portal through constant collection, implementation and validation of requirements. Likewise, the integrated data provision requires taking into account new datasets, the evolution and maintenance of existing data stewardship services, and data harmonization (e.g., using common standards for similar types of data).

This required a shared methodology to manage the interactions within the EPOS community and the establishment of an 'Interaction Team' to coordinate the collaborative work. Such common activities are organized with a systematic approach, inspired by the 'shape-up' method (Singer 2019) but reviewed to fit the research context (Figure 3). This choice was made after years of experience where the interactions were organized by adopting different methods, all inspired by traditional Software Development Life Cycle (SDLC) processes (Sommerville 2010), which require a detailed definition of the tasks to be done in advance (e.g., through an iteration in the Agile case (Martin & Micah 2006) and a meticulous resource (working time) allocation in advance.

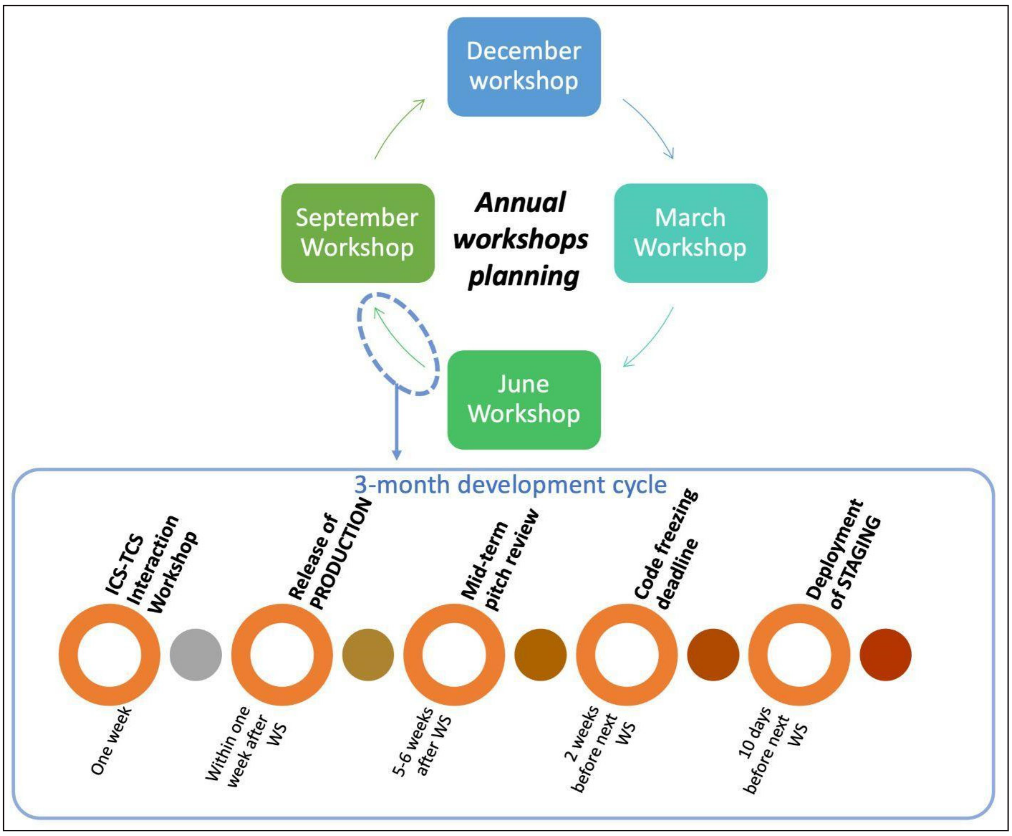


Figure 3 The 'shape-up' method reviewed to fit the research context. It envisages four workshops (WS) per year, followed by four development cycles.

EPOS successfully adopted the shape-up method, that somehow reverse-engineers the linear SDLC. In the shape up, indeed, on the basis of the deadlines and of the real available human resources, activities and tasks are defined, differently from the traditional SDLC methods where the starting points are the tasks. Shape up envisages the establishment of work cycles of a well-defined length at the end of which the expectation is to have a deliverable. This avoids slipping deadlines.

In the shape-up method, the work to be delivered is presented at the beginning of a work cycle in the form of a pitch (shaping phase). Pitches are small project planning documents, that include a) a description of the problem to be solved, b) the required efforts, c) a generic design with generic technical solutions, d) the risks affecting progress of the work (e.g., risk of having to refactor existing parts of the system to make it all work), d) NoGos, i.e., what is not going to be done in this work cycle, to avoid misunderstanding and to depict clear boundaries. Pitches are an exercise aimed at defining the boundary of the problem so that it can be addressed within the given deadlines and with the available resources.

In EPOS this method was implemented by establishing one-week *interaction workshops* arranged quarterly. In such workshops the developers, together with the scientists and the data providers, come together to report about previous work and to jointly discuss and agree on the pitches to be executed in the following work cycle. In the mid-term pitch review meeting, pitch leaders present progress to the IT Board, including representatives of the various key group of EPOS.[11] At this stage a contingency plan might be triggered. The entire work is coordinated by the Interactions Coordinator and aligned with the EPOS strategies and planning.

The implementation of the shape-up methodology shows that Federated Data Management requires also coordinating the community beyond the technical level, and that such community management activities require a clear sustainability schema (e.g., funding, clear roles, legal commitments) to enable all community members to be active parties in this process, with equal opportunities of influencing, steering and contributing to the federated Data provision.

EPOS DATA POLICY MANAGEMENT

Since the beginning, EPOS fostered the adoption of a common data policy applicable to the whole EPOS Delivery Framework. Indeed, in the case of large RIs like EPOS, the mere technical description of the data stewardship technical aspects (e.g., compliance with FAIR principles) is not sufficient, as in order to manage data within a well-defined legal framework, the RIs and the thematic communities have to converge towards a set of policies that take into account the governance of the distributed data provision system.

To design the EPOS Data Policy, a first exhaustive review of the existing national, European and international legislation was made during the EPOS preparatory phase (Kohler et al. 2017). In addition, most of the principles from the Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020 (EC 2017) were considered, despite EPOS not being part of the Open Research Data Pilot (ORD Pilot).[12]

In July 2018 the EPOS Data Policy was published as the official document implemented and endorsed by the whole EPOS community.[13] The EPOS Data Policy key principles are:

- to disseminate data and knowledge through Open Access.
- to make Data, Data products, Software and Services available in a timely manner, without undue delay and preferably free of charge taking in due account the need to differentiate between virtual and remote access and physical access.
- to follow the OECD principles (EC, 2017) for research data from public funding.
- to utilise a widely accepted community licensing scheme, i.e., Creative Commons.

The EPOS Data Policy acknowledges the ongoing work of the European Commission to foster the FAIR (Findable, Accessible, Interoperable, Reusable) principles for data access.

Once EPOS ERIC started to function as a legal entity in October 2018, the need to expand the EPOS Data Policy in view of the upcoming operational phase was recognised. To this end, a dedicated Working Group was established, that included different areas of expertise including legal, technical, policy, communication and management. As a result, the EPOS Digital Assets Management Policy was finalised, including relevant policies on Asset Management

[11] The EPOS IT Board includes : EPOS-ERIC IT-officer (IT-Board Chair), in charge of monitoring and managing the overall IT aspects of EPOS; ICS-C Technical coordinator, in charge of the ICS-C Integration system hosting and operation; Development coordinator, in charge of coordinating the software developments of the EPOS Data Portal; Interactions coordinator, managing the interactions team; representative of the data providers community, namely the Service Coordination Committee (SCC) Chair, guaranteeing that IT decisions are in line with the communities requirements; and the External Projects coordinator, in charge of keeping alignment between EPOS and other (pan)-European initiatives (e.g., European Open Science Cloud).

[12] In Horizon 2020 the Commission has launched a flexible pilot for open access to research data (ORD pilot). The pilot aims to improve and maximise access to and re-use of research data generated by Horizon 2020 projects, https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-dissemination_en.htm (accessed on 10 December 2022).

[13] EPOS Data policy is available at (https://www.epos-eu.org/sites/default/files/2020-12/EPOS%20DATA%20POLICY_July2018.pdf) (accessed on 10 December 2022).

(Curation, Provenance, Quality Assurance, Licensing, Citation/Acknowledgement) and Security (Physical Security, Authentication, Authorization, Disaster Recovery). The EPOS Digital Assets Management Policy should be read in conjunction with the EPOS ERIC Statutes[14] (C 2018/7011) and EPOS Data Policy, available on the EPOS website relevant sections.[15]

Thanks to this work, EPOS is regarded as a pioneering European Research Infrastructure for its established mechanisms for delivering policy documents. Indeed, in 2021 and 2022 EPOS was invited to training sessions organized by the ENVRI-FAIR project (Petzold et al., 2019), to present the work done by EPOS in this area to other environmental Research Infrastructures.

SUSTAINABILITY OF FEDERATED DATA MANAGEMENT

The EPOS research infrastructure emerged from the vision of stimulating state-of-the-art research and innovation by enabling *sustainable* and universal use and re-use of multidisciplinary data and products in solid Earth sciences.

This requires a shared vision on data management, and on the way to implement it given the implications at influencing the technical, governance, legal, scientific, and financial levels. In addition to the approaches described in previous sections, all contributing to ensuring sustainable access to data and services, EPOS developed specific tools to secure the sustainability of the infrastructure by addressing specific needs of the elements of the EPOS architecture (Figure 1).

First, concerning the Integrated Core Services Central Hub, EPOS ERIC has set up a multi-year partnership agreement (MYPA) with Hosting Organisations to regulate the technical, legal, financial and governance aspects of the hosting infrastructure (Contell et al. 2022). This agreement contains a technical annex that specifies the conditions to ensure continuous availability of the EPOS Data Portal, including resources and procedures for continuous and uninterrupted provision of the services, incident management, backup and restore. For instance, the amount of memory and CPUs allocated to the various components of the microservices based ICS-C system and the Key Performance Indicator (KPI) defining the availability of the system[16] (e.g., 99.3% availability with maximum 20 concurrent users per second, in line with the INSPIRE Guidelines for Download Services (Force 2010). It is therefore apparent that the Data Management at the integration infrastructure level requires the definition of robust specifications, agreed by the entire community, backed up by the adoption of legal agreements between the Hosting Organisations and EPOS ERIC to reinforce the sustainability of the infrastructure. Moreover, the multi-year partnership agreement was the concrete tool employed by EPOS ERIC to translate the commitment to host and operate the ICS-C declared by a set of countries members of the ERIC.

Second, to ensure the provision of data and products, the EPOS solution has been to establish legal agreements (Collaboration Agreements) (Contell et al. 2022) between EPOS ERIC and research organisations representing each thematic node (TCS). In particular, collaboration agreements cover: *TCS Governance and Coordination* and *TCS Data and Service provision*. Collaboration, the agreements with EPOS ERIC are established after a research community is recognized as TCS thematic community. The participation of research organizations in the TCS is governed by a Consortium Agreement (external to EPOS ERIC) to facilitate their establishment, administration, governance and operation. Consortium Agreements also stipulate the responsibility of research organisations in implementing the EPOS Data Policy. In this way, the key aspects of the Data Management Plans are shared by every research organisation delivering EPOS data and products. Collaboration Agreements also include specific tasks for the improvement of data access policies where gaps are identified, and to address the sustainability of data and services, of which data management is only one part. Collaboration Agreements also provide. Here, the list of services made available through EPOS is provided on a yearly basis, including the licensing status. These agreements also refer to the adoption of the EPOS Data Policy and Access Rules, which apply to the whole EPOS offer made by the TCS, whether accessible via the EPOS Data Portal or via specific thematic portals.

[14] https://www.epos-eu.org/epos-eric/documents (accessed on 10 December 2022).

[15] https://www.ics-c.epos-eu.org/data/search (accessed on 10 December 2022).

[16] 99.3% availability with maximum 20 concurrent users per second, in line with the INSPIRE Guidelines for Download Services (Force, 2010).

Third, concerning National Research Infrastructures (NRI), these are organisations outside the EPOS Delivery Framework, and EPOS ERIC does not have any legal link with them, except when NRIs are research organisations participating in the TCS. In this case, the adoption of practices aligned with EPOS by the NRIs is negotiated through the TCS, as part of the community building role of the TCS.

EPOS's multidimensional framework, with its TCS, ICS, and NRI components, illustrates the complexity and the evolving nature of distributed research infrastructures. This layered architecture accentuates the need for ad hoc agreements. While standardized agreements may offer a foundation, Schmiederer and Kuberek (2022) reminds us that there is not a universal blueprint for cooperation agreements. This entails that agreements crafted for the specific purpose are still a valid option in the case of distributed RIs.

These agreements not only clarify the intricate mutual relations between the different entities but also stipulate terms for data and service provision in line with the thematic DMPs and EPOS Policies. These policies, when paired with cooperation agreements, form the spine of EPOS's sustainable data management approach, crafting a clear path for both its current undertakings and future expansions.

In the EPOS's approach to Federated Data Management, all these elements are considered together to ensure a sustainable Research Data Management of distributed RIs.

4. DATA MANAGEMENT PLAN GUIDELINES FOR FEDERATED RESEARCH INFRASTRUCTURES

Horizon 2020 guidelines for FAIR Data Management (European Commission 2016) provide a framework for structuring data management plans and are organized in six different areas. The more recent guidelines for the Horizon Europe programme (European Commission 2021) improve the previous one by clarifying some concepts and by addressing the management not only of data but also of other research outputs, both digital and physical, ensuring alignment with FAIR principles for sharing and re-use. These guidelines are an effective tool, especially when integrated with other existing guidelines or tools like 'ERC Open Research Data and Data Management Plans' (ERC 2019), the already mentioned DMPonline tool (Donnelly, Jones & Pattenden-Fail 2010), the ARGOS tool which also allows generating machine actionable DMPs,[17] other tools dedicated to standard-based machine actionable data management (Cardoso, Castro & Miksa 2021), the Data Stewardship Wizard,[18] EasyDMP[19] and many others.

Comparative studies reviewing 14 different tools (Gajbe et al. 2021) suggest that there is a considerable overlap in terms of features addressed by the various DMP tools. However, those features concentrate mostly on technical aspects (e.g., security, storage, and backup), data related issues (e.g., FAIR, metadata, relationship among datasets), ethical dimension (e.g., roles and responsibilities) and to a limited extent also include Data Privacy. As such, most of the DMP related guidelines, tools and literature analysed so far, take into account in a very limited way some of the key dimensions that in the experience of EPOS are needed for properly managing Data in a distributed and federated environment.

With this in mind, we propose an extension of the Horizon Europe Guidelines for Data Management Plan (European Commission 2021), also including some of the key dimensions discussed in the previous sections. The extension is shown in Table I. The resulting 'Data Management Plan Guidelines for Federated Research Infrastructures' stems from the EPOS experience, but it is generic and can be potentially adopted by any Research Infrastructure that federates resources (e.g., data) that are provided in a distributed way.

In order to be coherent with the format used in the Horizon Europe guidelines, we hereby propose the additional components of an extended DMP by introducing some example questions addressing the key topics. The full description of these components is already discussed in the previous sections.

[17] https://argos.openaire.eu/splash/ (accessed on 10 December 2022).

[18] https://ds-wizard.org/ (accessed on 10 December 2022).

[19] https://easydmp.eudat.eu/ (accessed on 10 December 2022).

Table I **Components of the Data Management Plan Guidelines for Federated Research Infrastructures.** The table shows the Guidelines components suggested by the Horizon Europe Guidelines on FAIR Data Management, together with the proposed extension including additional components dedicated specifically to Data Management in the context of Distributed, Federated Research Infrastructures as stemming from the experience of EPOS.

DMP COMPONENT	DESCRIPTION
Components from Horizon Europe guidelines	
Data Summary	Brief description of the data, type of data and its coverage must be provided. Data formats, origins, total expected size of datasets and type of users should also be outlined
FAIR Data	Detailed description of how the FAIR principles are addressed and taken into account
Other Research Outputs	Plan for the management of other research outputs, digital or physical, that may be generated or re-used throughout the initiative in a FAIR way
Allocation of Resources	Description of resources to make data FAIR and to ensure its sustainability, as well as resources needed to deliver the plan
Data Security	Description of data recovery as well as secure storage and transfer of sensitive data
Ethical Aspects	Description of ethical or legal issues that can have an impact on data sharing
Other Issues	Additional information about other procedures for data management currently in place
Extended components for distributed, Federated Research Infrastructures	
Community Interaction	Description of the processes to manage the community of distributed resource providers in order to ensure federated provision of distributed resources
Governance	Description of the underlying governance at the basis of the federated resource provision
Policies	Addressing the policies that should be in place for supporting the governance and ensuring that resources (e.g., Datasets) combination, propagation and exchange is done appropriately and with well-defined responsibilities
Sustainability	Description of the legal, scientific and financial aspects that ensure the capacity to grant long-term (sustainable) access to data and products, and the capacity to do so under a common federated framework.

COMMUNITY INTERACTION

- Is the community structured in a federated way?
- If so, is it possible to identify responsible contacts for each of the federation ‘nodes’?
- Can the responsible contacts represent the resource provider node either for governance and for technical work? If not, does the community envisage having different individuals for representing technical and governance in the appropriate context?
- Is a procedure for fostering interaction among the technical contacts in place? If so, can it be described?
- What is the frequency of the interactions among the various node contacts?
- Are the results from the interactions monitored and summarised so that these can be shared with the entire community?

GOVERNANCE

- Are research organizations providing access to resources organized in consortia with well-defined agreements?
- Are research organizations providing access to resources compliant with the defined policies, including data management policies?
- Is there a governance in place for sustainably managing community-driven standards?
- Is a governance structure agreed between the research organizations providing access to federated resources for operating in a defined legal framework?
- Are there multi-annual agreements in place between the relevant parties defining the terms and conditions under which the Parties shall work in collaboration for their mutual benefit to enable the provision of data and services within the Federated Research Infrastructure?

POLICIES

- Are there policies on privacy, terms and conditions and cookies in place to behave correctly within law and prevent from legal litigations?
- Are there Digital Assets Management policies (e.g., provenance, quality assurance, licensing, security/authentication and authorization) in place to ensure that Data and Services are managed and used in ways that maximize public benefit following FAIR principles (Findability, Accessibility, Interoperability, and Reusability)?

SUSTAINABILITY

- Are there mechanisms in place to monitor the sustainability of data and service provision?
- Is the sustainability of data and service provision addressed from the federated governance framework?
- Is there long-term support available with respect to both human and financial resources?

5. DISCUSSION AND FUTURE WORK

The Data Management in a distributed Research Infrastructure is a multi-dimensional issue that requires addressing different aspects at the same time. We observe that, whereas filling in the questions with respect to sustainable FAIR data sharing in a standard DMP template often boils down to adding references to already firmly established certified components or standards (e.g., usage of Trusted Repositories Data Seal of Approval (Dillo & De Leeuw, no date), in the case of Federated RIs the full commitment to these requirements is actually delegated onto the deeper level of these components themselves. Therefore, the first and most relevant challenge is the ability to govern such a complex landscape taking into account a variety of issues, spanning from technical to financial, legal, ethical, governance and more. EPOS is addressing such a challenge through a set of procedures and rules that on one hand ensure that the management roles (e.g., Executive Director and Executive Coordination Office personnel) include a variety of expertise and skills which are required to manage the entire EPOS enterprise; on the other hand, the ERIC governance structure together with a set of implementing rules (Cocco et al. 2022b) and sustainability principles (including the financial aspect) ensure that the 'EPOS enterprise' is robust because it relies on the coordinated efforts of different teams that work together in a balanced and synergic way.

With reference to the Thematic data and services, the challenge is to keep a constant interaction across communities so to harmonize the efforts and have constant exchange of information, either in technical (e.g., mutual support for adopting common metadata standards) and in governance terms (e.g., adopting similar governance frameworks based on consortia). Of course, additional efforts are required to harmonize as much as possible the agreements – that are currently similar in the form but tailored to the specific community needs – and the data provision technical aspects. This latter point envisages the adoption of common IT procedures and tools to maximize the harmonization, for instance: common Persistent Identifiers; drafting of more clear and shared Data Management Plans based on the proposed Data Management Plan guidelines for Federated Research Infrastructures (DMP for RI) template presented above; common Authentication technologies (possibly based on OAuth2 or Open ID connect (Naik & Jenkins 2017) and interoperable Authorisation schema.

In perspective, the EPOS strategy envisages the inclusion of new Data, Data Products, Software and Services, and although it is clear that new resources need to be included into the Thematic Communities to be integrated – and go through a well-defined process including several steps of assessment – this raises the issue of the management of a continuously growing set of potential assets. This requires continuing and enhancing all harmonization and interaction activities, together with a plan for the sustainability of new resources. This should indeed be described through the proposed DMP for RI template.

As for the integrated data provision, this poses interesting challenges when it comes to the development of the so-called Integrated Core Services Distributed: these are intended as distributed resources like computational resources (HPC), remote environment for data analysis and processing (e.g., Jupyter Notebooks), visualization tools (e.g., enlighten-web) and reproducible analysis environments (Spinuso et al. 2022). This is an unexplored area, where first steps have been done on the technical side, but the sustainability and the governance dimension still need to be explored and established.

In terms of Hosting Infrastructure, the challenge is the need of constant monitoring its availability either in terms of e-Infrastructure underpinning the EPOS Data Portal, and in terms of continuous monitoring of the data sources availability, which are committed to provide access to the assets in a robust way with KPIs for availability (e.g., 99% availability over one year). Steps forward have been done with the definition of the 'aspirational' KPIs derived from the INSPIRE framework, and with the implementation of GUI oriented indicators (e.g., user alerts when datasets are not available), but still much work is required to make the entire data provision more robust. KPIs related to the availability of services should indeed also be part of the DMP in a Federated RI environment whenever the RI is required to be Operational by means of legal commitments.

The interaction between the key actors through the adapted shape-up method is a significant achievement in the management of the data provision. However, it needs to be constantly monitored, revised and updated to match the evolution of the community, to ensure that within each thematic community there is an appropriate exchange of information, and to ensure that meetings are not too time consuming for the individuals involved in this process. A well-defined balance between in-person, remote meetings and frequency of the workshops has therefore to be constantly reframed. Although the adapted shape-up methodology has proved to work well in EPOS, it clashes somehow with the linear SDLC-like habits of IT developers, so constant monitoring of its appropriate adoption is required.

As for the policies, a set of core policy documents together with guidelines about how to implement them was elaborated. This poses however the challenge of ensuring that thematic communities and underpinning NRIs are factually convergent on the implementation of these policies, which are intended to become real practices in the day-to-day work.

On the sustainability side, interactions with NRIs indicate that EPOS could play a role in supporting the sustainability of data provision by granting more visibility to NRIs. In this way, NRIs could increase their impact and benefit, for instance, from more competitive funds.

Additional work in the framework of another EPOS initiative (EPOS Sustainability Phase project[20]), point to risks in the financial sustainability of thematic services elaborated for EPOS. This is expected particularly in the case of new services not existing before EPOS. Here, the sustainability of data management is directly linked to the capacity to deliver EPOS data in the long term, and as such needs to be described in detail in the DMP.

Discussions with other Research Infrastructures with similar features, i.e., federating distributed resource providers, for instance in the framework of cluster initiatives like ENVRI-FAIR, suggest that the issues encountered in the framework of EPOS (and discussed in the current work) are common to other RIs. In this sense, the proposed *Data Management Plan guidelines for federated Research Infrastructures,* which can be potentially the basis for a shared approach to Data Management in the Federated RI context, needs to be further developed, validated and discussed also in domains different from solid Earth Science. This approach may indeed improve the Federated Data Management and support the RI in the difficult task of Managing Data in such complex landscapes.

6. CONCLUSIONS

In this paper we address the topic of Data Management in distributed, Federated Research Infrastructures. In particular, we present and discuss the experience of EPOS (European Plate Observing System). Due to its structure, which relies on a federated approach aggregating more than 250 data and software services into one single Data Portal, EPOS adopted a specific Federated Data Management approach which includes the elements from the Horizon 2020 guidelines and goes beyond to address dimensions peculiar to federated environments.

Data Management Plans (DMPs) are indeed powerful tools enabling the management of data. In the last decade, DMPs have evolved considerably into online tools (e.g., DMPonline) and techniques based on standards for enabling machines to read and interpret DMPs. However, when adopted in the context of Federated Research Infrastructures, these DMP templates and tools present critical gaps. In the case of EPOS, to fill such gaps, five dimensions were considered for an appropriate Federated Data Management.

The first dimension is the management of the data sources which required a significant work in terms of harmonization of: a) the specific DMPs of the various service providers, b) the establishment of agreements with EPOS ERIC for data provision, c) technical harmonization with particular reference to the FAIR principles.

[20] The EPOS SP – European Plate Observing System Sustainability Phase project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 871121.

The second dimension addresses the management of the data integration infrastructure. It required the construction of a dedicated integration service (ICS-C) accessible through the EPOS Data Portal. The Portal requires securing the hosting resources through agreements with the Hosting Organizations on the basis of a Technical Annex. At the same time, a set of practices and guidelines was agreed with the thematic data providers for managing the technical details of the integration.

The third dimension is related to the management of the interactions across the key actors of the data provision, namely the thematic communities and the integration IT team. Such interaction enables data providers, developers', and operators' teams to guide the evolution of the EPOS Data Portal and of the thematic data provision services through constant collection, implementation and validation of requirements and of development practices. Various attempts, following standard linear-like Software Development Life Cycles which first define the tasks and then allocate the resources, proved of limited effectiveness. As such, EPOS decided to adopt a method inspired by the shape-up framework, which is characterized by an approach that begins with the definition of resources and then defines, in cascade, the tasks that can be realistically performed. This approach demonstrated to be oriented to delivery, and is currently managed through iterative cycles, based upon plenary workshops with a three-month frequency, where all the key actors get together for revising the work done, for defining tasks for the next cycle and for discussing urgent issues.

The fourth dimension underpinning the entire data provision, and only partially addressed by the Horizon Europe template, is related to sustainability. This dimension includes, of course, the financial aspect, but in the case of distributed Research Infrastructures like EPOS it required the setting of governance principles and legal arrangements to ensure that data and products would be properly managed in a sustainable way. The scientific aspect is also part of sustainability.

Finally, policies were elaborated for ensuring that the datasets integration and propagation is done appropriately, with well-defined responsibilities, aiming at providing open, free and easy access to data.

Based on these dimensions, we propose an extension of the Horizon 2020 Guidelines for Data Management template, which includes four additional elements, namely community interaction, governance, policies, sustainability. The community interaction element addresses the processes to manage the data providers to ensure federated provision of distributed resources. The governance element is needed to manage the components of the architecture at the basis of the federated resource provision. The policy element is introduced for supporting the governance and ensuring that dataset combination and propagation is done appropriately and with clear responsibilities. Finally, sustainability, that addresses the legal, scientific, and financial aspects, is also included as it ensures the capacity to grant long-term (sustainable) access to data and products, and the capacity to do so under a common federated framework.

The proposed *Data Management Plan guidelines for Federated Research Infrastructures,* further developed, validated, and discussed in different domains, can potentially be the basis for a common approach, adopted by Federated Research Infrastructures out of EPOS. This approach may be discussed also at European Commission level for providing a more comprehensive and fit for purpose Data Management Plan guidelines for Federated Research Infrastructures.

ACKNOWLEDGEMENTS

The Authors acknowledge: all the EPOS thematic communities, the EPOS IT Development Team, the Hosting Organization Team, the EPOS ERIC staff, the EPOS graphic designer Barbara Angioni, Dr. Massimo Cocco for his endless efforts in making EPOS real, the EU Commission that supported EPOS through several projects (EPOS-PP – GA no 262229, EPOS-IP – GA no 676564, EPOS-SP – GA no 871121).

COMPETING INTERESTS

The authors have no competing interests to declare.

AUTHOR CONTRIBUTIONS

Daniele Bailo conceived the original concept of the paper and drafted the core structure of the paper. *Rossana Paciello* supported drafting the structure of the paper and contributed to the Management of the Data Integration Infrastructure section. *Jan Michalek* contributed to the Management of Thematic Communities for Integrated Data Provision section. *Daniela Mercurio* contributed to the drafting of the Data Policy Management section with the support of *Giovanna Maracchia*. *Agata Sangianantoni* and *Kauzar Saleh Contell* contributed to the drafting of the Sustainability of Federated Data Management section. *Otto Lange* contributed to the conception of the key concept of the paper and in particular to the Data Management Plan Guidelines for Federated Research Infrastructures section. *Kuvvet Atakan* and *Keith G Jeffery* contributed to the drafting of the EPOS Federated Approach section. *Lilli Freda* supervised the drafting of the paper, performed a general revision and contributed to the conclusion section.

AUTHOR AFFILIATIONS

Daniele Bailo orcid.org/0000-0003-0695-4406
Istituto Nazionale di Geofisica e Vulcanologia, Rome, Italy

Rossana Paciello orcid.org/0000-0002-6975-1991
Istituto Nazionale di Geofisica e Vulcanologia, Rome, Italy

Jan Michalek orcid.org/0000-0002-8057-7541
Universitetet i Bergen (UiB) 5020 Bergen, Norway

Daniela Mercurio
European Plate Observing System, EPOS ERIC, Rome, Italy

Agata Sangianantoni orcid.org/0000-0003-2564-4032
Istituto Nazionale di Geofisica e Vulcanologia, Rome, Italy

Kauzar Saleh Contell
European Plate Observing System, EPOS ERIC, Rome, Italy

Otto Lange orcid.org/0000-0003-3560-988X
Utrecht University, Utrecht, The Netherlands

Giovanna Maracchia orcid.org/0009-0004-9273-6097
Istituto Nazionale di Geofisica e Vulcanologia, Rome, Italy

Kuvvet Atakan orcid.org/0000-0002-0297-2060
Universitetet i Bergen (UiB) 5020 Bergen, Norway

Keith G. Jeffery orcid.org/0000-0003-4053-7825
Keith G Jeffery Consultants, United Kingdom

Carmela Freda orcid.org/0000-0002-2320-8096
European Plate Observing System, EPOS ERIC, Rome, Italy

REFERENCES

Atakan, K, et al. 2019. *EPOS-IP Project – Deliverable 6.7 – Second Delivery of Data Management Plan* (September).

Bailo, D, Jeffery, KG, et al. 2022. Data integration and FAIR data management in Solid Earth Science. *Annals of Geophysics*, 65(2): DM210. DOI: https://doi.org/10.4401/ag-8742

Bailo, D, Paciello, R, et al. 2022. Integrated access to multidisciplinary data through semantically interoperable services in a metadata-driven platform for Solid Earth Science, *MTSR 2022*. In: *16th International Conference on Metadata and Semantics Research*. London, 7th–11th November 2022, pp. 1–14.

Best, MMR, et al. 2016. *The EMSO-ERIC Pan-European Consortium: Data benefits and lessons learned as the legal entity forms*. 5(3): 8–15. DOI: https://doi.org/10.4031/MTSJ.50.3.13

Bristol University. 2014. EC Horizon 2020 Pilot on Open Research, pp. 1–7.

Cardoso, J, Castro, LJ and **Miksa, T.** 2021. Interconnecting systems using machine-actionable data management plans – hackathon report. *Data Science Journal*, 20(1): 1–11. DOI: https://doi.org/10.5334/dsj-2021-035

Cocco, M, et al. 2022a. The EPOS Research Infrastructure: a federated approach to integrate solid Earth science data and services. *Annals of Geophysics*, 65(2): 1–15. DOI: https://doi.org/10.4401/ag-8756

Cocco, M, et al. 2022b. The EPOS Research Infrastructure: a federated approach to integrate solid Earth science data and services. *Annals of Geophysics*, 65(2): DM208. DOI: https://doi.org/10.4401/ag-8756

Contell, KS, et al. 2022. Long-term sustainability of a distributed RI: the EPOS case. *Annals of Geophysics*, 65(2): 1–16. DOI: https://doi.org/10.4401/ag-8786

Dillo, I and **De Leeuw, L.** no date. *Data seal of approval: Certification for sustainable and trusted data repositories 2 DATA SEAL OF APPROVAL Data Seal of Approval*. Available at: http://www.datasealofapproval.org [Accessed: 26 November 2018].

Donnelly, M, Jones, S and **Pattenden-Fail, JW.** 2010. DMP online: A demonstration of the digital curation centre's web-based tool for creating, maintaining and exporting data management plans. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 6273 LNCS(October): pp. 530–533. DOI: https://doi.org/10.1007/978-3-642-15464-5_74

EPOS IP WP6 and **WP7 Teams.** 2015. EPOS ICS-TCS Integration Guidelines – Handbook for TCS integration: Level-2. *Zenodo*. DOI: https://doi.org/10.5281/zenodo.34666

European Commission. 2013 (December). *Guidelines on Data Management in Horizon 2020*. Version 1, p. 6. Available at: http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf.

European Commission. 2016 (July). *H2020 Programme. Guidelines on FAIR Data Management in Horizon 2020*, Version 3, p. 12. Available at: https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf.

European Commission. 2021 (May). Horizon Europe. Data Management Plan Template. Version 1.0. Available at: https://ec.europa.eu/info/funding-tenders/opportunities/portal/screen/how-to-participate/reference-documents;programCode=HORIZON, under: Templates & forms – Project reporting templates – Data management plan.

European Commission (EC). 2017 (March). Guidelines to the rules on Open Access to Scientific Publications and Open Access to research data in Horizon 2020. Available at: https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf.

European Research Council (ERC). 2019 (July). Open Research Data and Data Management Plans: Information for ERC grantees, p. 19. Available at: https://erc.europa.eu/sites/default/files/document/file/ERC_info_document-Open_Research_Data_and_Data_Management_Plans.pdf.

European Strategy Forum on Research Infrastructures (ESFRI). 2021. Roadmap & Strategy Report on Research Infrastructures. DOI: https://doi.org/10.25607/OBP-1861

Force, I. 2010 (Oct). Infrastructure for Spatial Information in Europe Technical Guidance for the implementation of INSPIRE Download Services, (1088), pp. 1–89.

Gajbe, SB, et al. 2021. Evaluation and analysis of Data Management Plan tools: A parametric approach. *Information Processing and Management*, 58(3): 102480. DOI: https://doi.org/10.1016/j.ipm.2020.102480

Jones, S, et al. 2020. Data management planning: How requirements and solutions are beginning to converge. *Data Intelligence*, 2(1–2): 208–219. DOI: https://doi.org/10.1162/dint_a_00043

Karasti, H, et al. 2018. *Little Data, Big Data, No Data? Data Management in the Era of Research Infrastructures*.

Kohler, E, et al. 2017. *D4.2 Guidelines for implementing the EPOS data policy and access rules*. *European Commission*, 18: 1–26. Available at: https://ec.europa.eu/research/participants/documents/downloadPublic?documentIds=080166e5b190e31a&appId=PPGMS.

Marti, M, et al. 2022. Addressing the challenges of making data, products, and services accessible: An EPOS perspective. *Annals of Geophysics*, 65(2): DM212. DOI: https://doi.org/10.4401/ag-8746

Martin, RC and **Micah, M.** 2006. *Agile Principles, Patterns, and Practices in C#*.

Michener, WK. 2015. Ten simple rules for creating a good Data Management Plan. *PLoS Computational Biology*, 11(10): 1–9. DOI: https://doi.org/10.1371/journal.pcbi.1004525

Miksa, T, et al. 2018. Defining requirements for machine-actionable Data Management Plans. *Zenodo*. Available at: https://zenodo.org/record/1266211#.XOb2oYhKiUk.

Naik, N and **Jenkins, P.** 2017. Securing digital identities in the cloud by selecting an apposite Federated Identity Management from SAML, OAuth and OpenID Connect. In: *International Conference on Research Challenges in Information Science*. Briton, UK on 10–12 May 2017, pp. 163–174. DOI: https://doi.org/10.1109/RCIS.2017.7956534

Paciello, R, et al. 2022. EPOS-DCAT-AP 2.0 – State of play on the Application Profile for metadata exchange in the EPOS RI. DOI: https://doi.org/10.1007/978-3-031-39141-5_21

Petzold, A, et al. 2019. ENVRI-fair-interoperable environmental fair data and services for society, innovation and research. In *IEEE 15th International Conference on eScience*. San Diego, CA: eScience on 24 Jul 2019, pp. 277–280. DOI: https://doi.org/10.1109/eScience.2019.00038

Sallans, A and **Donnelly, M.** 2012. DMP online and DMPTool: Different strategies towards a shared goal. *International Journal of Digital Curation*, 7(2): 123–129. DOI: https://doi.org/10.2218/ijdc.v7i2.235

Schmiederer, S and **Kuberek, M.** 2022. Forschungsdaten-Policy für Forschungsprojekte im Spannungsfeld zwischen Kooperationsvertrag und Datenmanagementplan: Untersuchung und Abgrenzung zentraler Dokumente in Forschungsvorhaben. *Bausteine Forschungsdatenmanagement*, 2). DOI: https://doi.org/10.17192/bfdm.2022.2.8446

Singer, R. 2019. *Shape up stop running in circles and ship work that matters*. Basecamp.com, V1.8, pp. 1–135. Available at: https://basecamp.com/shapeup.

Sommerville, I. 2010. *Software engineering. Software Engineering*. DOI: https://doi.org/10.1111/j.1365-2362.2005.01463.x

Spinuso, A, et al. 2022. SWIRRL. Managing provenance-aware and reproducible workspaces. *Data Intelligence*, 4(2): 243–258. DOI: https://doi.org/10.1162/dint_a_00129

Swauger, S. 2015. ‘DMPTool. *The Charleston Advisor*, 16(3): 12–15. DOI: https://doi.org/10.5260/chara.16.3.12

Vermeulen, A, et al. 2015. ICOS Carbon Portal Progress Report 2014 & 2015.

Wilkinson, MD, et al. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3: 160018. DOI: https://doi.org/10.1038/sdata.2016.18

Williams, M, Bagwell, J and **Nahm Zozus, M.** 2017. Data management plans, the missing perspective. *Journal of Biomedical Informatics*, 71: 130–142. DOI: https://doi.org/10.1016/j.jbi.2017.05.004

TO CITE THIS ARTICLE:
Bailo, D, Paciello, R, Michalek, J, Mercurio, D, Sangianantoni, A, Contell, KS, Lange, O, Maracchia, G, Atakan, K, Jeffery, KG and Freda, C. 2024. Data Management in Distributed, Federated Research Infrastructures: The Case of EPOS. *Data Science Journal*, 23: 5, pp. 1–17. DOI: https://doi.org/10.5334/dsj-2024-005

Submitted: 15 December 2022
Accepted: 19 December 2023
Published: 24 January 2024

Data Science Journal is a peer-reviewed open access journal published by Ubiquity Press.