Towards Operational Research Infrastructures with FAIR Data and Services

Zhiming Zhao[1(✉)], Keith Jeffery[2], Markus Stocker[3,4], Malcolm Atkinson[5], and Andreas Petzold[6]

[1] Multiscale Networked Systems, University of Amsterdam, 1098XH Amsterdam, The Netherlands
z.zhao@uva.nl
[2] Keith G Jeffery Consultants, Faringdon, UK
keith.jeffery@keithgjefferyconsultants.co.uk
[3] TIB Leibniz Information Centre for Science and Technology, Hannover, Germany
markus.stocker@tib.eu
[4] MARUM Center for Marine Environmental Sciences, PANGAEA Data Publisher for Earth & Environmental Science, Leobener Strasse 8, 28359 Bremen, Germany
[5] University of Edinburgh, Edinburgh, UK
malcolm.atkinson@ed.ac.uk
[6] Forschungszentrum Jülich GmbH, Jülich, Germany
a.petzold@fz-juelich.de

Abstract. Environmental research infrastructures aim to provide scientists with facilities, resources and services to enable scientists to effectively perform advanced research. When addressing societal challenges such as climate change and pollution, scientists usually need data, models and methods from different domains to tackle the complexity of the complete environmental system. Research infrastructures are thus required to enable all data, including services, products, and virtual research environments is FAIR for research communities: Findable, Accessible, Interoperable and Reusable. In this last chapter, we conclude and identify future challenges in research infrastructure operation, user support, interoperability, and future evolution.

Keywords: Research infrastructure · Virtual research environment · System-level science

1 Introduction

Natural and anthropogenic factors lead to environmental changes on all scales from local to global. Environmental data provides the scientific basis for analysing the physical, biological, and economic processes in the earth system, which are affecting all sectors of society as well as wildlife and biodiversity. Such data-related activities can be highlighted in several scenarios drawn from research communities, as shown in Fig. 1:

Z. Zhao and M. Hellström (Eds.): Towards Interoperable Research Infrastructures for Environmental and Earth Sciences, LNCS 12003, pp. 360–372, 2020.
https://doi.org/10.1007/978-3-030-52829-4_20

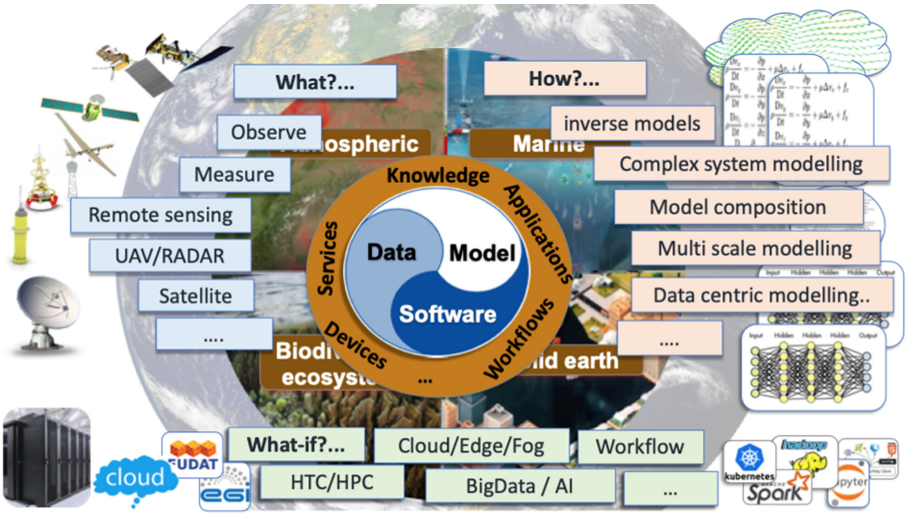

Fig. 1. Some typical research activities in the environmental RI communities. (Zhao Z. presentation in the final ENVRIplus event in Brussel, June, 2019.)

1. **Observing** the phenomena of the environmental and earth system, via distributed sensors, monitoring network or human observers [1]. Such observations are often guided by specific conceptual models of the parameters related to the earth system, or connected with the experimentations in the laboratory (e.g. rock mechanics) or in the fields (agricultural studies).
2. **Modelling** the behaviour of the environmental systems, understanding their evolution, and investigating the causality among different events by scientists [2]. These models can be developed based on physical models, e.g. the Navier–Stokes equation for modelling fluid dynamics [3], machine learning methods like neural networks [4], or combinations thereof [5].
3. **Applying** existing assets from observations, simulations, and earlier experiments to complex data-centric workflows to explore the solution space of hypotheses or discover the consequences of different conditions [6]. Scientific workflows, e.g. [7] or Jupyter notebook [8] are used to integrate different processes in the data pipeline, which may involve big data processing platforms, e.g. Spark [9], across different infrastructures, e.g. Cloud and HPC clusters [10].

As important facilities to enable scientists to perform advanced research in environmental and earth sciences, environmental research infrastructures aim to make their digital assets, including data, models and software Findable, Accessible, Interoperable and Reusable (**FAIR**). In a broad sense, all the **application workflows, the sensors** that obtain data**, the operational services** that manage those assets**, and all high-level knowledge** derived from those assets, collectively constitute valuable material for user communities to conduct scientific research with, as indicated in Fig. 1. However, the tools and infrastructures to manage, document, provide, find, access, and use all such assets

are still underdeveloped owing to a combination of data complexity and increasingly large data volumes.

We have seen the large collection of research infrastructures proposed and developed by different communities. Figure 2 provides a basic landscape of those infrastructures, which are scattered across four subdomains: the atmospheric domain, e.g. IAGOS and ACTRIS, the marine domain e.g. Euro-Argo and EMSO, the solid earth domain, e.g. EPOS, and the ecosystem and biosphere domains, e.g. AnaEE and DISCCO. Some of them cross multiple domain boundaries, e.g. ICOS and Lifewatch.

Fig. 2. The landscape of ENVRI research infrastructures across domains. (Image source: https://envri.eu/communications/)

In this last chapter, we will first give a short summary of the main topics discussed in the book, and then look at the next phase of research infrastructure evolution: new challenges that RI may face after they are developed and deployed.

2 ENVRI: Development Activities at the Cluster Level

The development of an environmental RI is often driven by the interests of a specific domain, and many RIs are funded via separate research projects. But the bottom-up development paradigm of the RIs does not come with a naturally embedded interoperability concern for guiding the evolution of those different projects. Therefore, dedicated cluster projects are funded since 2011 to specifically inspect common problems that

these environmental RIs face, to recommend reusable solutions to developers of RIs, and to tackle the interoperability challenge among RIs. Figure 3 shows the three cluster projects funded for ENVRI RIs: ENVRI, ENVRIplus [11] and ENVRI-FAIR [19]. The work presented in this book is mainly based on the activities carried out in the second project.

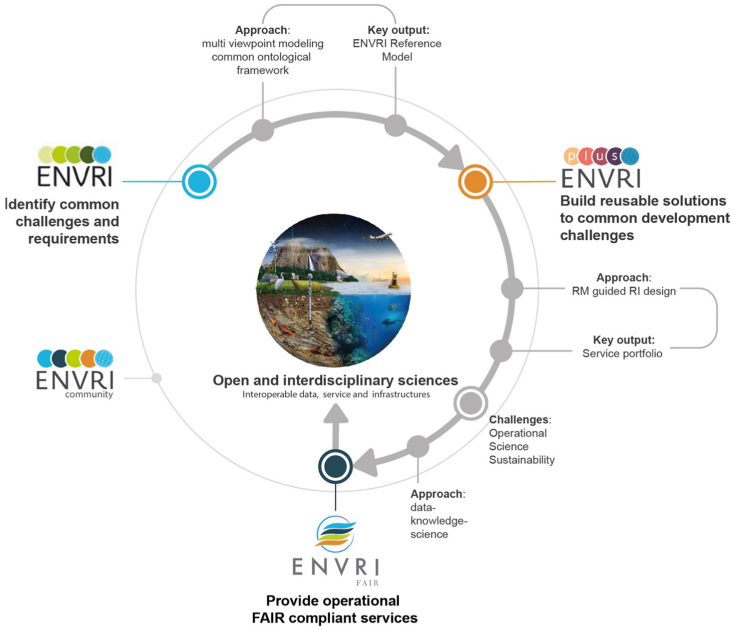

Fig. 3. The activities in the cluster of ENVRI research infrastructures.

1. In the FP7 ENVRI project (between 2011 and 2014), we analysed the initial design requirements, architecture design, and the existing assets of participating research infrastructures[1]. We abstracted a common vocabulary (ENVRI RM V1) for describing data management activities and architecture of research infrastructure.
2. Using the initial ENVRI RM, we analysed the design requirements from more than 20 research infrastructures[2] in the follow-up project H2020 ENVRIplus and applied a reference model guided approach to design and prototype six prioritised six common operations. During the practice, the ENVRI RM has also been refined as V2.

[1] FP7 ENVRI project contains six ESFRI projects: EMSO, Euro-Argo, ICOS, LifeWatch, EISCAT_3D and EPOS.

[2] H2020 ENVRIplus project contains 21 RIs: ACTRIS, ANAEE, EISCAT_3D, ELIXIER, EMBRC, EMSO, EPOS, ESONET, Euro-Argo, EUROFLEETS, EUROGOOS, FIXO3, IAGOS, ICOS, INTERACT, IS-ENES, JERICO, LifeWatch, LTER, SeaDataNet2 and SIOS.

3. In the H2020 ENVRI-FAIR project, we focus on the operational challenges, in particular, the FAIRness of the assets[3]. By the moment we finalise the book, the ENVRI-FAIR project just finished its initial self-assessment of FAIRness.

In this section, we will summarise our development activities during the past years via a number of highlights.

2.1 A Common Vocabulary for Describing Data Management

The ENVRI Reference Model[4] (ENVRI RM) was created at the beginning of the ENVRI project (the first cluster project) as a common ontological framework to enhance the information sharing among different research infrastructures [12] (see Chapter 4). The development of the ENVRI RM started from the data management lifecycle of research infrastructures in ENVRI community and abstracted five common phases: acquisition, curation, publishing, processing and use. Following a multi-view approach provided by the ODP (Open Distributed Processing) model, the ENVRI RM team abstracts the key vocabularies for describing communities, behaviour, data flow management, service interfaces and architectures patterns from ENVRI research infrastructures. The initial ENVRI RM has been refined with a big set of RIs in the ENVRIplus project. The ontological representation of ENVRI RM has also been created for the machine-readable specification (Fig. 4).

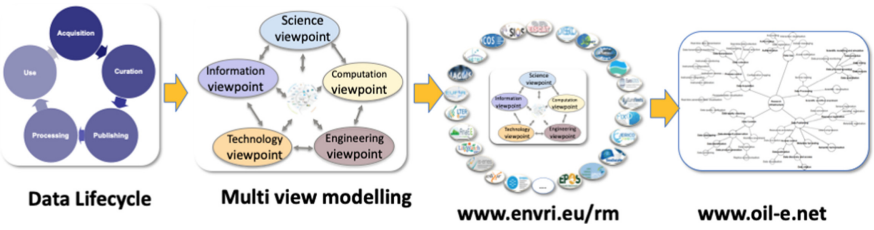

Fig. 4. The development approach for the common ENVRI vocabulary.

2.2 Reference Model Guided Engineering

Using the common ontological framework, we analysed the requirements for research infrastructure and common challenges in the ENVRIplus project. After several iterations, we highlighted the six key common challenges: identification/citation, data processing, infrastructure optimisation, curation, cataloguing, and provenance, as discussed in Chap. 5.

The reference model enables the development team of common operations (i.e. data for science theme in ENVRIplus) effectively interact with the developers from different

[3] H2020 ENVRI-FAIR project contains 13 RIs: ACTRIS, ANAEE, DANUBIUS-RI, DiSSCo, EISCAT_3D, EPOS, EMSO, Euro-Argo, IAGOS, ICOS, LifeWatch, eLTER and SIOS.

[4] http://envri.eu/rm.

research infrastructures, and from the eInfrastructures to 1) analyse requirements, 2) review technologies and gaps, 3) design solutions to the common problem and 4) validate the prototypes via use cases. The details of the approach are discussed in Chapter 5.

The development team developed or recommended the key technologies for tackling the common problem identified from the research infrastructure:

1. Reference model and relevant training materials (in Chapter 4);
2. Ontological representation of the reference model (in Chapter 6);
3. Data curation services and recommendations (in Chapter 7)
4. Data cataloguing services (in Chapter 8)
5. Data identification services and citation recommendation (in Chapter 9)
6. Data processing framework and technologies (in Chapter 10)
7. Virtual infrastructure for data-centric sciences (in Chapter 11)
8. Data provenance services and recommendation (in Chapter 12)
9. Metadata and semantic linking (in Chapter 13)
10. Authentication, Authorisation, and Accounting (in Chapter 14)
11. Virtual Research Environment (in Chapter 15)

During the ENVRIplus project, the key results [11] have been documented and collected as a portfolio.

2.3 Use Case-Based Community Engagement

To best engage user communities in the development, ENVRI follows an **Agile development methodology**. Selected use cases follow a continuous procedure for accepting and reviewing proposals, prioritising specific use case projects, setting up agile use case projects, monitoring progress and exploiting results.

Based on the size and scope of individual cases, we identified three types of use cases, as shown in Fig. 5.

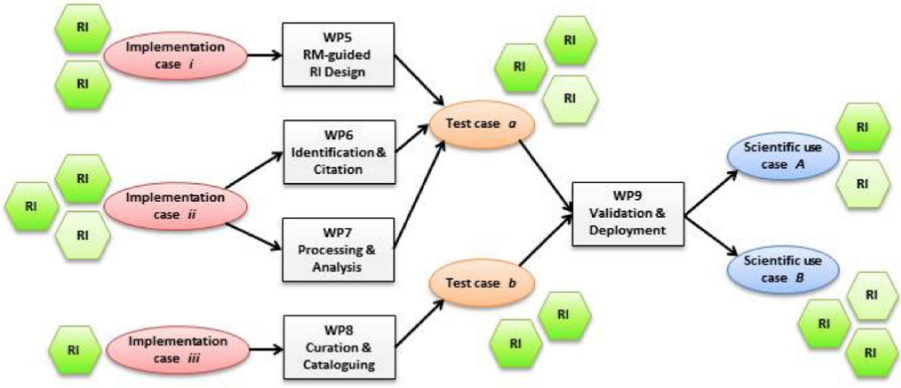

Fig. 5. Associations between science, implementation and test cases with core ENVRIplus activities. (http://www.envriplus.eu/wp-content/uploads/2015/08/D9.1-Service-deployment-in-computing-and-internal-e-Infrastructures.pdf)

1. **Science cases** often have clearly defined scientific problems, and require a big development effort for technical components, e.g. integrating data or services from different infrastructures.
2. **Test cases** focus on specific problem research infrastructures are facing, and often require the implementation of 1 or two critical components in the case. It can be implemented within typically a half year.
3. **Implementation cases** focus on specific technologies (e.g. customisation, integration or minor modification), and most of the components involved in the case are already available. The test cases can be implemented within typically 2–3 months.

In practice, the outputs of the implementation cases provide useful input for implementation cases and finally contribute to the development of science cases. Each use cases project often has members from different task teams and execute in parallel with the project task teams. In this way, the developers of the common data services participate in one or more agile case projects and closely collaborate with members from the research infrastructure communities.

Within each use case team, regular telcos are organised. By reviewing the progress, the developers can adapt the action points to meet the changing demands from the RI communities. In this book, three typical use cases have been presented in Chapter 16, 17 and 18.

2.4 A Community Knowledge Base

By the end of ENVRIplus, most of the RIs in the cluster have either finished their preparation phase or their implementation phase and are ready for final implementation or operation respectively. Collecting information about the RI's implementation status and the tools and technologies that they were using (including software, standards and vocabularies) was deemed vital for coordinating collaboration and identifying key commonalities. To this end, an ENVRI Knowledge Base is prototyped in the latter stages of ENVRIplus, for further development during the ENVRI-FAIR successor project, with the goal of using the ENVRI RM as a basis for modelling the active state of different ENVRI RIs. The knowledge base, along with the semantic technology applied in its creation, is discussed more fully in Chapter 6 of this book (Fig. 6).

The RI status, including architecture and available data management services, the service portfolio, and the FAIRness self-assessment (performed in the ENVRI-FAIR project) have been ingested in the knowledge base. The details of the knowledge base have been discussed in [13] and Chapter 6.

2.5 Lessons Learned

During the lifetime of ENVRI and ENVRIplus, there has been a challenge in interactions between specialists of the generic IT technologies, and the software developers of the research infrastructures. Although the generic IT specialists can clearly see the technical problems and gaps; since those specialists are not embedded in the development

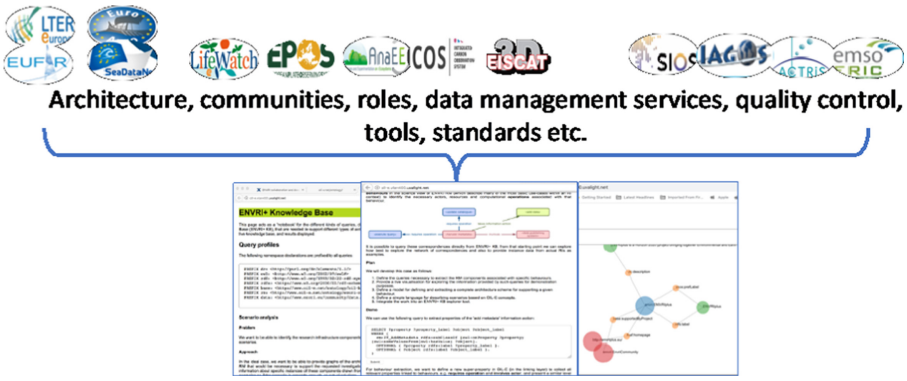

Fig. 6. Knowledge for sharing best practices.

context of each individual RI, the proposed solutions often miss matches the development priorities and the user practices of an individual RI. The interactions are thus often time-consuming.

The classical waterfall model of software engineering did not work in this context. The interactions between generic IT specialists and RI developers need to be spiral and iterative. Figure 7 shows how the other key highlights during the interactions. Besides what we have discussed in this section, two summer schools have been organised for transferring the technical knowledge to the RI developers, by the time when the book finishes. The key output has also been exploited to the third cluster project ENVRI-FAIR for the further development of the RI data management services, to make them FAIR compliant and operational.

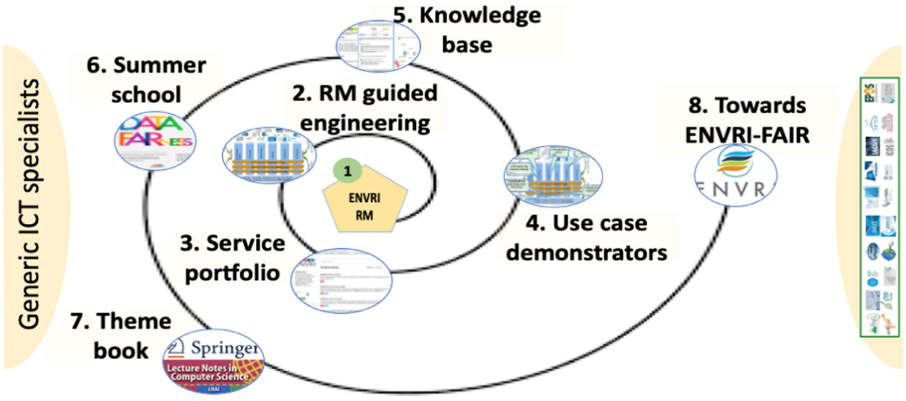

Fig. 7. Key highlights between Specialists and the RI communities.

3 Looking at the Next Steps

Upon becoming fully operational, each RI has to face an increasingly large number of users, and an increasing number of different user scenarios involving their digital assets. In this context, the developers and operators of the RI services have to face several challenges, as discussed below.

3.1 Towards European Open Science Cloud (EOSC)

ENVRI-FAIR is the connection of the ENVRI Cluster to the European Open Science Cloud (EOSC). The overarching goal is that by the end of the project, all participating RIs will have built a set of FAIR data services which enhance the efficiency and productivity of researchers, support innovation, enable data- and knowledge-based decision making and connect the ENVRI Cluster to EOSC and its core services. This goal is reached by (1) well-defined community policies and standards addressing all steps of the data lifecycle, aligned with the wider European policies and international developments; (2) each participating RI having sustainable, transparent and auditable data services for each step of data lifecycle, compliant with the FAIR principles; (3) focusing the proposed work into the implementation of prototypes for testing pre-production services at each RI, with the catalogue of prepared services defined for each RI independently depending on the maturity of the involved RIs; and (4) exposing the complete set of thematic data services and tools provided by the ENVRI cluster under the EOSC catalogue of services.

3.2 Operational Challenges

When operated as online services, RIs collaborate naturally as part of a service ecosystem, wherein each RI has to serve users from much bigger user communities than its own community, e.g. when enabling users to perform system-level science. Moreover, RIs are often part of a global network of infrastructures focused on the same subject, besides being a member of the European ENVRI RI community, e.g. Euro-Argo in the global network Argo and eLTER in the LTER federation. A RI, therefore, needs to optimise its operational models with consideration of the practices of the wider network or cluster. A number of challenges can, therefore, be highlighted:

1. Defining effective operational models which can help RIs exploit the existing e-infrastructures contributing to EOSC as well as their own computing and instrumentation infrastructure. A RI will balance disruption against assured benefits as it engages to maximise resources and gains interoperability with other infrastructures.
2. Authenticating and authorising users from different communities to use shared resources, and accounting for the usage of the data, services and underlying e-infrastructure within a framework of trust, security and privacy.
3. Allowing technical coordination across RIs through appropriate interfaces; this entails adopting interfaces for supporting shared VREs [14, 15], contributing rich (FAIR-compliant) metadata to community catalogues.
4. Ensuring the performance and quality of service and user experience required by scientists, in a manner that scales with the user base and data assets.

5. Effectively provisioning RI resources, including data and tools offered by RIs and services delivering underlying data infrastructure, to serve a broad range of demands from research developers, service managers, engineers, and researchers themselves.
6. Integrating EOSC with Fog/Edge computing scenarios and IoT (Internet of Things); some RIs have extensive sensor networks and technology which needs to be connected to the broader e-infrastructure.

3.3 Science Challenges

Many ENVRIplus stakeholders have stated that the community of environmental research infrastructures should be closely involved in EOSC developments. The ENVRIplus approach may help shape the EOSC ecosystem. More importantly, significant parts of the ENVRI community stand to benefit from EOSC. This transition will be incremental, as relevant services become available, affordable and sustainable, and when they combine well with current investments and agreed practices. ENVRIplus stakeholders are in a good position to bring in crucial views and development actions to support open science in the whole research process.

As the basis for open science, FAIR (or more appropriately FAIR+R where the additional R is reproducible) data, services and other relevant resources require not only incentives for sharing and exploiting data on the part of data producers and users, but also the development of effective technologies and standards that will enable RIs to achieve connectivity and interoperability of their data and services at any stage in the data management lifecycle.

To enable system-level and interdisciplinary science, future RIs have to face the following challenges:

1. Enabling interdisciplinary research activities to meet environmental research goals and societal challenges; not only sharing research data and software assets from different RIs, but also co-developing and using methodologies and models drawing expertise from multiple domains within and outside of environmental science [17].
2. Ensuring that the data and resources needed by scientists follow FAIR principles; this means the services, methods and metadata to make these assets FAIR.
3. Supporting user-specified and steered data processing, and automated workflows. For example, many user requests result in a workflow to download one or more selected datasets. As services local to data become well-supported then users develop and use more complex workflows involving multiple datasets, software components, computing resources and even sensors with processing partitioned and local to the data assets. The generation of workflows from user requests and their optimal deployment will grow in importance for environmental research.
4. Recording and providing provenance information for user assessment of relevance and quality of an asset, auditing, and reproduction.
5. Reusing the data and knowledge from different RIs effectively; this requires effective data and knowledge mining tools and a cohesive support knowledge infrastructure.

6. Providing support for data-intensive, compute-intensive and urgent data analysis and simulation. Frequently, complex workflows using such simulations need to inter-work between HPC (High-Performance Computing) and HTC (High Throughput Computing) platforms.
7. Providing support for working across multiple e-infrastructure environments within EOSC and beyond (e.g. DataOne in the USA [16]). RI workflows may utilise EOSC and other e-Is (including sensor networks) together and the interface should allow 'plug and play'.
8. RIs may be involved in activities with RIs on other continents and so may need to access e-Is in those other continents (and vice-versa) through appropriate gateways.

3.4 Sustainability Challenges

In the previous chapter, sustainability was specifically discussed. The operators of the RIs have to face several challenges to keep their services sustainable, including:

1. Providing sustainable business models that service data contributors, service developers, researchers, innovation makers and other payers into EOSC can use to ensure their continued participation [18].
2. Providing sustainable data management and stewardship, including the curation, long-term preservation and access of assets (information and software including associated libraries and operational environment).
3. Providing sustainable technical decisions, including standards and interfaces, so that they fit with the evolution of the digital ecosystem and operational models of RIs.
4. Providing sustainable system architecture and accompanying engineering to meet demands for scaling technical solutions for larger numbers of users.
5. Choosing effective underlying infrastructure for provisioning RIs and deploying services to achieve sustainable service quality and reliability avoiding 'lock-in' to any particular set of e-Is.
6. Educating RI researchers, managers, developers, curators and other actors on how to utilise EOSC through their RI appropriately.

4 Concluding Remarks

The ENVRIplus project ended in July 2019. Although the main content of this book is based on the output of the ENVRIplus project, the community effort put into ENVRI continues into the ENVRI-FAIR project and other collaborative and interoperability initiatives. We hope this book provides a valuable summary of the knowledge we developed in the project and enhances the transfer of knowledge to the development and user communities of the ENVRI and other scientific infrastructures.

Acknowledgements. This work was supported by the European Union's Horizon 2020 research and innovation programme via the ENVRIplus project under grant agreement No 654182.

References

1. Tanhua, T., et al.: Ocean FAIR data services. Front. Mar. Sci. **6**, 440 (2019). https://doi.org/10.3389/fmars.2019.00440
2. Brunner, D., et al.: Comparison of four inverse modelling systems applied to the estimation of HFC-125, HFC-134a, and SF6; emissions over Europe. Atmos. Chem. Phys. **17**, 10651–10674 (2017). https://doi.org/10.5194/acp-17-10651-2017
3. Woodring, J., Petersen, M., Schmeiber, A., Patchett, J., Ahrens, J., Hagen, H.: In situ eddy analysis in a high-resolution ocean climate model. IEEE Trans. Visual. Comput. Graphics. **22**, 857–866 (2016). https://doi.org/10.1109/TVCG.2015.2467411
4. Kurth, T., et al.: Exascale deep learning for climate analytics. In: SC18: International Conference for High-Performance Computing, Networking, Storage and Analysis, pp. 649–660. IEEE, Dallas (2018). https://doi.org/10.1109/SC.2018.00054
5. Kutz, J.N.: Deep learning in fluid dynamics. J. Fluid Mech. **814**, 1–4 (2017). https://doi.org/10.1017/jfm.2016.803
6. Hey, T., Tansley, S., Tolle, K. (eds.): The Fourth Paradigm: Data-Intensive Scientific Discovery. Microsoft Research, Albuquerque (2009)
7. Atkinson, M., Gesing, S., Montagnat, J., Taylor, I.: Scientific workflows: past, present and future. Future Gener. Comput. Syst. **75**, 216–227 (2017). https://doi.org/10.1016/j.future.2017.05.041
8. Prathanrat, P., Polprasert, C.: Performance prediction of Jupyter notebook in JupyterHub using machine learning. In: 2018 International Conference on Intelligent Informatics and Biomedical Sciences (ICIIBMS), pp. 157–162. IEEE, Bangkok (2018). https://doi.org/10.1109/ICIIBMS.2018.8550030
9. Stocia, I.: Conquering big data with spark. In: 2015 IEEE International Conference on Big Data (Big Data). p. 3. IEEE, Santa Clara (2015). https://doi.org/10.1109/BigData.2015.7363734
10. Evans, K., et al.: Dynamically reconfigurable workflows for time-critical applications. In: Proceedings of the 10th Workshop on Workflows in Support of Large-Scale Science - WORKS 2015, pp. 1–10. ACM Press, Austin (2015). https://doi.org/10.1145/2822332.2822339
11. Ari, A., et al.: Final ENVRIplus project report, (2019). Zenodo https://zenodo.org/record/3517905
12. Martin, P., et al.: Open information linking for environmental research infrastructures. In: 2015 IEEE 11th International Conference on e-Science, pp. 513–520. IEEE, Munich (2015). https://doi.org/10.1109/eScience.2015.66
13. Zhao, Z., et al.: Knowledge-as-a-service: a community knowledge base for research infrastructures in environmental and earth sciences. In: 2019 IEEE World Congress on Services (SERVICES), pp. 127–132. IEEE, Milan (2019). https://doi.org/10.1109/SERVICES.2019.00041
14. Martin, P., Remy, L., Theodoridou, M., Jeffery, K., Sbarra, M., Zhao, Z.: Mapping heterogeneous research infrastructure metadata into a unified catalogue for use in a generic virtual research environment. Future Gener. Comput. Syst. **101**, 1–13 (2019). https://doi.org/10.1016/j.future.2019.05.076
15. Hu, Y., et al.: Deadline-aware deployment for time critical applications in clouds. In: Rivera, F.F., Pena, T.F., Cabaleiro, J.C. (eds.) Euro-Par 2017. LNCS, vol. 10417, pp. 345–357. Springer, Cham (2017). https://doi.org/10.1007/978-3-319-64203-1_25
16. Sandusky, R.J.: Computational provenance: DataONE and implications for cultural heritage institutions. In: 2016 IEEE International Conference on Big Data (Big Data), pp. 3266–3271. IEEE, Washington DC (2016). https://doi.org/10.1109/BigData.2016.7840984

17. Casale, G., et al.: Current and future challenges of software engineering for services and applications. CloudForward (2016). http://dx.doi.org/10.1016/j.procs.2016.08.278
18. Petzold, A., Asmi, A.: ENVRI-FAIR EOSC Position Paper (2020). Zenodo http://doi.org/10.5281/zenodo.3666806
19. Petzold, A., et al.: ENVRI-FAIR - interoperable environmental FAIR data and services for society, innovation and research. In: 2019 15th International Conference on eScience (eScience), pp. 277–280. IEEE, San Diego (2019). https://doi.org/10.1109/escience.2019.00038, https://zenodo.org/record/3462816