Common Challenges and Requirements

Barbara Magagna[1(✉)], Paul Martin[2], Abraham Nieva de la Hidalga[3], Malcolm Atkinson[4], and Zhiming Zhao[2]

[1] Environment Agency Austria, Vienna, Austria
barbara.magagna@umweltbundesamt.at
[2] Multiscale Networked Systems, University of Amsterdam, 1098XH Amsterdam, The Netherlands
pwmartin.research@gmail.com, z.zhao@uva.nl
[3] Cardiff University, Cardiff, UK
nievadelahidalgaa@cardiff.ac.uk
[4] Edinburgh University, Edinburgh, UK
mpa@staffmail.ed.ac.uk

Abstract. Research infrastructures available for researchers in environmental and Earth science are diverse and highly distributed; dedicated research infrastructures exist for atmospheric science, marine science, solid Earth science, biodiversity research, and more. These infrastructures aggregate and curate key research datasets and provide consolidated data services for a target research community, but they also often overlap in scope and ambition, sharing data sources, sometimes even sites, using similar standards, and ultimately all contributing data that will be essential to addressing the societal challenges that face environmental research today. Thus, while their diversity poses a problem for open science and multidisciplinary research, their commonalities mean that they often face similar technical problems and consequently have common requirements when addressing the implementation of best practices in curation, cataloguing, identification and citation, and other related core topics for data science.

In this chapter, we review the requirements gathering performed in the context of the cluster of European environmental and Earth science research infrastructures participating in the ENVRI community, and survey the common challenges identified from that requirements gathering process.

Keywords: Requirements · Environmental science · Research infrastructure

1 Introduction

Today's societal challenges such as hazard mitigation and sustainable resource provision need new interdisciplinary approaches pooling, resources, insights, data methods and models. The transformation of the way to analyse natural phenomena through the advent of digital instruments and the intensive use of computers also known as the "Fourth Paradigm" [1] came along with a wealth of data that poses immense challenges in how to handle and exploit it. This transition is common to all environmental and Earth sciences

Z. Zhao and M. Hellström (Eds.): Towards Interoperable Research Infrastructures for Environmental and Earth Sciences, LNCS 12003, pp. 30–57, 2020.
https://doi.org/10.1007/978-3-030-52829-4_3

and many research communities are engaged in generating and exploiting that data. To optimise the working practices and the platforms that support the data pipelines from distributed data sensors to storage, use, presentation and analysis, it is crucial to study the condition of research infrastructure and systems currently available.

In the context of the ENVRI community, this becomes even more important as all the RI initiatives share common challenges, both in their construction and operation. Although each RI has its specific ICT strategy, there are potential benefits from defining, designing, and developing shared solutions for common problems. The use of common solutions also fosters strategies for sharing data or reducing the barriers to data interoperability. The ENVRI community thus encourages and facilitates joint work to develop synergies, learn from each other, harmonise the RI service landscape, and share best practices. Since 2011, the ENVRI community has been supported by multiple projects aiming to improve the collaboration within the European Research Area and beyond. Thus, requirements elicitation has a long tradition within the ENVRI community and with each ENVRI supporting project this effort has been repeated with the aim to record the last status of the involved RIs in the light of emerging technologies and newly pressing challenges.

The first ENVRI project (FP7 funded, 2011–2014) analysed the design of six participating ESFRI environmental research infrastructures (ICOS, Euro-Argo, EISCAT_3D, LifeWatch, EPOS, and EMSO) to identify common computational characteristics and requirements. The ISO/IEC standard Open Distributed Processing (ODP), a multi-viewpoint conceptual framework for building distributed systems, was used as a common platform for interpretation and discussion to ensure a unified understanding. The outcome of this endeavour proved to be helpful for the understanding of strengths and weaknesses in the outline and planned developments of the RIs.

Within the time frame of the ENVRIplus project (2015–2019) [2], the status of involved RIs (which now numbered over 20) along the dimensions of data, users, software services and resources were re-analysed. The aim was to identify commonalities and differences between RIs and to point to the state-of-the-art of RI technologies. The characterisation of RIs under a standard documentation method with vocabulary defined in the ENVRI Reference Model (see chapter 4) allowed comparison and discussion leading to best practice and consistent development plans for RI improvements.

For the requirements study, a standard method for describing all relevant ICT aspects needed to provide the facilities and capabilities required by RI researchers was defined. This process led to the identification of seven specific topics identified in the project [2]:

- **Identification and citation.** Mechanisms to provide and retrieve durable references to data objects and collections of data objects.
- **Curation.** Processes to assure the availability and quality of data over the long term.
- **Cataloguing.** The construction of discoverable and searchable indexes of datasets, processes, software and other resources made available by RIs.
- **Processing.** Computational transformations of data, including but not restricted to processing and selection of raw data close to instruments, signal processing, analysis of data for quality assurance (QA) purposes, simulation runs with a subsequent comparison with observations, and statistical analyses.

- **Provenance.** Processes to record information about how data, code and working practices were created and were transformed into their current form.
- **Optimisation.** Methods for improving the quality of service offered to researchers as data and processing requirements increase, focusing mainly on the underlying movement and processing of data.
- **Community support.** Addressing all aspects of the use of resources by researchers and their relationships with resource providers.

The requirements of the ENVRI RIs were distilled and assessed in the context of these seven topics; from this assessment, a number of general requirements can be seen that remain broadly applicable even as the research landscape continues to evolve, incorporating the concepts of FAIR [3] data and European Open Science Clouds [4]. Part of this chapter is drawn from the project requirement analysis deliverable [5].

2 Requirements Collection in ENVRI

The ENVRI community brings together environmental and Earth science RIs, projects, networks and technical specialists with the joint ambition to create a holistic, coherent, interdisciplinary and interoperable cluster of environmental research infrastructures across Europe (and beyond). To do this, the ENVRI community brings together roughly four different environmental domains: *atmospheric*, *marine*, *biosphere/ecosystem* and *solid earth*. By working together, the idea is to capitalise on the progress made in various disciplines and strengthen interoperability amongst RIs and domains to better support more interdisciplinary research of the sort needed to address modern environmental grand challenges.

Table 1 gives an overview of the RIs that participated in the requirements gathering activity during the ENVRIplus project (2015–2019). The table indicates their organisation type, domain and involved data life cycles, as defined by the ENVRI RM. These RIs are typically composed of distributed entities (e.g. data generators, data processors and data sharers) and thus federations of often diverse autonomous organisations. These organisations have established roles, cultures, working practices and resources, and they often play roles in different federations of international infrastructures. The organisations' roles thus must remain unperturbed.

The original analysis of requirements gathered by the ENVRIplus project is provided by Atkinson et al. [5], but in this chapter we review some of the details that remain most pertinent to continued RI development in the environmental sciences and beyond. Most RIs focus on one domain, but several address multiple domains, typically with a common factor in mind (e.g. carbon observation or observations based on a specific region of interest, such as the Arctic). We look at each domain in turn.

2.1 Atmospheric Domain

Atmospheric science RIs typically seek to provide scientists and other user groups with free and open access to high-quality data about atmospheric aerosols, clouds, and trace

Table 1. Overview of RIs contributing to requirements gathering.

RI	Type of RI	Domain	Data Lifecycle
ACTRIS	Distributed	Atmospheric	Use to publishing
AnaEE	Distributed	Ecosystem	Curation to processing
EISCAT-3D	Single RI, multi-site	Atmospheric	Use to publishing
ELIXIR	Distributed	Ecosystem	Acquisition to publishing
EMBRC	Distributed	Marine, Ecosystem	Use to publishing
EMSO	Single RI, multi-site	Marine	Acquisition to publishing
EPOS	Distributed	Solid Earth	Acquisition to publishing
Euro-Argo	Distributed	Marine	Production to publishing
EuroGOOS	Distributed	Marine	Production to publishing
FixO3	Distributed	Marine	Acquisition to publishing
IAGOS	Distributed	Atmospheric	Acquisition to processing
ICOS	Distributed	Atmospheric, Marine, Ecosystem	Acquisition to publishing
INTERACT	Distributed	Ecosystem	Acquisition to publishing
IS-ENES2	Virtual	Multi-domain	Acquisition to publishing
LTER	Distributed	Ecosystem	Use to publishing
SeaDataNet	Virtual	Marine	Acquisition to publishing
SIOS	Distributed	Multi-domain	Publishing

gases from coordinated long-term observations, complemented with access to innovative and mature data products, together with tools for quality assurance, data analysis and research. Data are typically gathered by a mix of both ground and in-air sensors (e.g. aircraft-mounted), with different focuses on particular types of observation (e.g. greenhouse gases) or specific regions (e.g. the arctic, or the upper atmosphere).

RIs in ENVRI focusing on the atmospheric domain include: ACTRIS (Aerosols, Clouds, and Trace gases Research Infrastructure), which integrates European ground-based stations for long-term observations of aerosols, clouds and short-lived gases; EISCAT_3D, which operates next-generation incoherent scatter radar systems for observation of the middle and upper atmosphere, ionosphere and near-Earth objects such as meteoroids; ICOS, which provide long-term observations to understand the behaviour of the global carbon cycle and greenhouse gas emissions and concentrations and IAGOS (In-service Aircraft for a Global Observing System), which implements and operates a global observation system for atmospheric composition by deploying autonomous instruments aboard commercial passenger aircraft.

ACTRIS has requirements for 1) improving interoperability so as to make their data as accessible and understandable as possible to others; 2) understanding best practices when researchers need to discover data, particularly given the experiences of other RIs; 3) planning and managing the activity of sensors; 4) developing understanding of how

instruments work in extreme conditions; and 5) improving the capabilities of small sensors. ACTRIS is therefore concerned with such issues as data visualisation, data provisioning and interoperability between data centre nodes.

EISCAT_3D will contribute to our understanding of the near-Earth space environment for decades to come; whereas the domain of research is common with existing environmental RIs, EISCAT_3D will differ from most of those in its modes of operation and volumes of data. It is planned that at least 2 petabytes (PB) of data per year will be selected for long-term curation and archival. EISCAT_3D will thus present both high throughput computing (HTC) and big data analysis and curation challenges similar to those encountered in the particle physics and astrophysics communities. A common and increasingly important issue is citability and reusability of the data, as embodied in the FAIR concept. To keep track of data use is also a requirement from many funding bodies, including the member science councils and institutes of EISCAT. Therefore EISCAT is preparing to adopt a common scheme of persistent IDs for Digital Objects, such as fractional Digital Object Identifiers (DOI). EISCAT thus has requirements for workflow specification, data access with search and visualisation, interoperability with other RIs and instruments via virtual observatories.

ICOS aims to provide effective access to a single and coherent data set to enable research into the multi-scale analysis of greenhouse gas emissions, sinks and the processes that determine them. ICOS is concerned with 1) metadata curation, 2) data object identification and citation and 3) data collection and management of provenance information.

IAGOS provides freely accessible data for users in science and policy, including air quality forecasting, verification of CO_2 emissions and Kyoto monitoring, numerical weather prediction, and validation of satellite products. IAGOS expected through its participation in ENVRIplus to improve data discovery, metadata standardisation, interoperability and citation and DOI management. It also expected ENVRIplus to provide services for, citation, cataloguing and provenance.

2.2 Marine Domain

Marine domain RIs are concerned with observations and measurements of marine environments both near the coast and out to sea, near the oceans' surface and on the seafloor, and everywhere in-between [6]. As for atmospheric domain RIs, different RIs may focus on specific types of measurement or on specific sub-environments. Marine RIs have been observed to have the greatest maturity overall in terms of established data standards and curation practices for certain classes of dataset when compared to the other environmental science domains.

RI projects in ENVRI focused on the marine domain include: EMBRC (European Marine Biological Resource Centre), a distributed European RI which is set up to become the major RI for marine biological research, covering everything from basic biology, marine model organisms, biomedical applications, biotechnological applications, environmental data and ecology; EMSO (the European multidisciplinary seafloor & water column observatory), a large-scale RI for integrating data gathered from a range of ocean observatories; Euro-Argo, the European contribution to the Argo initiative, which provides data service based on a word-wide deployment of robotic floats

in the ocean; EuroGOOS (European Global Ocean Observing System), an international Not-for-Profit organisation promoting operational oceanography, i.e., the real-time use of oceanographic information; FixO3 (Fixed Open Ocean Observatory network), a research project that integrates oceanographic data gathered from a number of ocean observatories and provides open access to that data to academic researchers; and SeaDataNet, a Pan-European infrastructure for ocean & marine data management, which provides on-line integrated databases of standardised quality.

In what concerns data, the role of EMBRC is to generate and make it available. It does not usually do any analysis of those data, unless it is contracted to do so. Data is usually generated through sensors in the site in the sea or samples that are collected and then measured in the lab. EMBRC aimed to achieve several objectives through participation in ENVRIplus: establishing collaborations with the environmental community, which would benefit from their environmental and ecological data; developing and learning about new standards and best practices in terms of standards; developing new standards within INSPIRE, which can be used for other datasets; exploring new data workflows, which make use of marine biological and ecological data; and networking with other RIs. EMBRC requires common standards and workflows, harmonisation of data between labs, backup systems, maintenance of software and their integration into a single platform.

A goal of EMSO is to harmonise data curation and access, while averting the tendency for individual institutions to revert to idiosyncratic working practices after any particular harmonisation project has finished. There is a notable overlap between EMSO and FixO3 data (i.e., some FixO3 data is provided within the EMSO infrastructure). EMSO would like to obtain with the help of ENVRIplus better mechanisms for ensuring harmonisation of datasets across their distributed networks. Heterogeneous data formats increase the effort that researchers must invest to cross discipline boundaries and to compose data from multiple sources. Improved search is also desirable; currently expert knowledge is required, for example to be able to easily discover data stored in the MyOcean environment. Furthermore, EMSO is investigating collaborations with data processing infrastructures such as EGI for providing resources for infrastructure-side data processing. EMSO requires data interoperability across distributed networks and data search.

Like EMSO, FixO3 requires better mechanisms for ensuring harmonisation of datasets across their distributed networks. Heterogeneous data formats make life difficult for researchers. Improved search is also desirable; currently expert knowledge is required, for example to be able to easily discover data stored in the MyOcean environment.

FixO3 also requires harmonisation of data formats and protocols across their distributed networks, as well as harmonisation of data curation and access.

The Euro-Argo research infrastructure comprises a central facility and distributed national facilities. Euro-Argo aims at developing the capacity to procure and deploy and monitor 250 floats per year and ensure that all the data can be processed and delivered to users (both in real-time and delayed-mode). Euro-Argo sought within ENVRIplus to design and pioneer access to and use of a cloud infrastructure with services close to European research data to deliver data subscription services. Users would provide their

criteria: time, spatial, parameter, data mode, update period for delivery (daily, monthly, yearly, near real-time).

EuroGOOS strives to improve the coordination between their different member research institutes. Another important role of EuroGOOS is that of facilitating access to data for their community. Through participation in ENVRIplus, EuroGOOS valued: learning about other European RIs and getting inspiration from them for deciding on the general objectives and services that they could provide at European level; from a technological perspective, getting recommendations about the design of their common data system, including formats or data platforms and data treatments; getting inspiration from other RIs about ways to distribute the data to end users using applications which are more focused in this respect; and improved data assimilation.

Regarding SeaDataNet, the on-line access to in-situ data, metadata and products is provided through a unique portal interconnecting the interoperable node platforms constituted by the SeaDataNet data centres. SeaDataNet wanted to enhance the cross-community expertise on observation networks, requirements support and data management expertise by participating in ENVRIplus. More specifically, SeaDataNet needs technical support for cross-community (ocean, solid earth and atmosphere) visibility of information provided by SeaDataNet (platforms, metadata, datasets and vocabulary services), as well as expertise on interoperability services and standards.

SeaDataNet requires data policy to involve data providers in the publication of their own datasets.

2.3 Ecosystem Domain

Ecosystem or biodiversity RIs focus on the study and monitoring of biological ecosystems both large and small with the objectives of both providing accurate and in-depth information about the condition and spread of various species throughout the Earth and their interactions, and developing models that can predict how ecosystems will evolve under various conditions or may have evolved in the past.

RI projects in ENVRI focused on the ecosystem/biodiversity domain include: AnaEE (Analysis and Experimentation on Ecosystems), which focuses on providing innovative and integrated experimentation services for ecosystem research; ELIXIR, a European infrastructure for biological information that unites Europe's leading life-science organisations in managing and safeguarding the massive amounts of data being generated every day by publicly funded research; INTERACT (International Network for Terrestrial Research and Monitoring in the Arctic), a circum-arctic network of 76 terrestrial field stations in various northern nations for identifying, understanding, predicting and responding to diverse environmental changes; and LTER (Long-Term Ecosystem Research), a long-standing alliance of researchers and research sites invested in better understanding the structure, functions, and long-term response of ecosystems to environmental, societal and economic drivers.

AnaEE aims to provide excellent platforms with clear accessibility conditions and service descriptions, and a clear offering to researchers. The gathering of information in a common portal should help with this. Experiences gathered from the construction and operation of other platforms would be helpful to shape this development. Within the context of ENVRIplus, AnaEE was particularly interested in participating in the work on

identification and citation and on cataloguing, as these were of fairly immediate concern to their infrastructure, and consequently, it was useful to synchronise their approach with other RIs. Processing is of some interest as well, in particular the interoperability between models and data, and the quality control of data produced by platforms.

ELIXIR will provide the facilities necessary for life-science researchers—from bench biologists to chemo-informaticians—to make the most of our rapidly growing store of information about living systems, which is the foundation on which our understanding of life is built. By participating in ENVRIplus, ELIXIR aimed to establish a closer collaboration with other environmental RIs and improve their access to life science data. An enhanced interaction, a better insight into data structures and relevant data standards widely adopted across environmental RIs can facilitate an effective evaluation of areas of collaboration for development of new tools, services and training. Ultimately, this can lead to better interoperability and discoverability of environmental and life science data by users across atmospheric, marine, solid earth and biosphere domains.

INTERACT is keen on working on homogenisation with other infrastructures. The most important bilateral benefits of NordGIS (the INTERACT geographical metadata information system[1]) versus ENVRIplus are the broad European standards exposed to NordGIS, as well as the grass-root requirements exposed to ENVRIplus. INTERACT is open for new interactive solutions, and recognises that standards on how to turn primary data into data products suitable for OPEN dissemination need to be adopted. INTERACT needs to move into the realm of handling actual data concerning active field stations.

Due to the fragmented character of LTER Europe, harmonised data documentation, real-time availability of data as well as harmonisation of data and data flows are the overarching goals. As of the most recent review, LTER Europe is developing a Data Integration Portal (DIP, e.g. including a time series viewer) and is working on the integration of common data repositories into their workflow system (including metadata documentation with LTER Europe DEIMS[2]). Therefore, based on the common reference model, the outputs of ENVRIplus can provide development advice on those matters, which would be appreciated by LTER.

2.4 Solid Earth Domain

RIs in the solid earth domain are concerned with the study of seismology, volcanology, geodesy and other research disciplines focused on the Earth beneath our feet. The RI project focused on the solid earth domain in ENVRI is EPOS (European Plate Observing System), a long-term plan for the integration of Research Infrastructures for solid Earth science in Europe. Its main aim is to integrate communities to make scientific discovery in the domain of solid Earth science, integrating existing (and future) advanced European facilities into a single, distributed, sustainable infrastructure (the EPOS Core Services).

EPOS will enable the Earth Science community to make a significant step forward by developing new concepts and tools for accurate, durable, and sustainable answers to societal questions concerning geo-hazards and those geodynamic phenomena (including

[1] http://www.nordgis.org/.
[2] https://deims.org/.

geo-resources) relevant to the environment and human welfare. EPOS wanted advice from ENVRIplus to improve the Interoperable AAAI (Authentication, Authorisation, Accounting and Identification) system (federated & distributed), taking already existing software and make it available and scalable across communities.

2.5 Cross-Domain Concerns

Not all RIs neatly fit their activities into a particular domain as defined above. Many RIs address cross-cutting environmental concerns such as greenhouse gas emissions, and so integrate facilities and data sources with very different characteristics that might simultaneously contribute to more dedicated RIs as well.

RI projects participating in the ENVRI community that have a notable cross-domain focus include: ICOS (Integrated Carbon Observation System), an RI providing the long-term observations required to understand the present state and predict future behaviour of the global carbon cycle and greenhouse gas emissions and concentrations; IS-ENES2, contributing to the European Network for Earth System Modelling; and SIOS (Svalbard Integrated Earth Observing System), an integral Earth Observing System built to better understand the on-going and future climate changes in the Arctic.

The objectives of ICOS are to provide effective access to a single and coherent dataset to facilitate research into the multi-scale analysis of greenhouse gas emissions, sinks and the processes that determine them, and to provide information, which is profound for research and for the understanding of regional budgets of greenhouse gas sources and sinks, their human and natural drivers, and the controlling mechanisms. This requires insight into the interaction between atmospheric, marine and ecosystem datasets and services in particular.

IS-ENES encompasses climate models and their environment tools, model data and the interface of the climate modelling community with high-performance computing, in particular the European RI PRACE. Its requirements were mainly collected from the climate-modelling community, two data-dissemination systems (ESGF for project run time; LTA as long-term archiving), CMIP5 as climate modelling data project (2010–2015) and CMIP6 (2016–2021). By participating in ENVRIplus IS-ENES2 expected to obtain a better understanding of interdisciplinary use cases and end-user requirements, as well as advice for data catalogues to compare their model data with other data (e.g. observations). IS-ENES2 requires sharing of best practices as new nodes are integrated into the RI federation; keeping data near to processing; handling of volume and distribution of data; replication and versioning; and providing related information for data products (e.g. provenance, annotations and metadata).

Currently, SIOS is building a distributed data management system called SIOS Knowledge Centre, to develop methods for how observational networks are to be designed and implemented. The centre will lay the foundation for better-coordinated services for the international research community with respect to access to infrastructure, data and knowledge management, sharing of data, logistics, training and education.

2.6 Overall Requirements

It can be seen that there is substantial variability by RI and by topic. For every RI, a significant effort was made to develop communication and obtain information about requirements for all relevant topics. In some cases the RI was mature, in the sense that the RI or those involved in the work being done within the RI had been active in the particular domain for a significant number of years; the marine RIs that are already sharing data, such as Euro-Argo and SeaDataNet are good examples. Such maturity leads to an appreciation of the complexities and significance of various requirements. In other cases, the RI concerned was in a consortium of interacting, often global, related communities that share data and hence appreciate many of the issues; EPOS is one such example. For such RIs, it was possible to gather good input on virtually every topic. For all of the RIs, contact was made and information was gathered for at least the general requirements. In some cases, an RI deemed their interests were already covered by another RI known to be similar with which they worked closely.

The variation between topics is also a manifestation of variation in RI maturity. Topics such as identification and citation, cataloguing and processing are all encountered at the early stages of an RI's development and at early stages of the data lifecycle. Conversely, the value of topics such as curation and provenance become much more apparent after running a data gathering and sharing campaign for long periods or from being involved in the later stages of the data lifecycle. Optimisation is an extreme example of this effect; only when production and a large community of users demand more resources than an RI can afford does optimisation become a priority; before that the focus is on delivering the breadth of functionality users require and promoting adoption by a substantive community.

The overall conclusion would be that there are many opportunities for benefit from sharing ideas, methods and technologies between RIs, that there is much potential for using their data in combination and that there is a general need for awareness raising and training. However, these high-level consistencies have to be treated with great care; there are many lower level details where differences are significant. Continued work is needed to tease out those differences that are fundamentally important and which are coincidental results from the different paths that RI projects have already taken to date. Fundamental differences need recognition and support with well-developed methods for linking across them founded on scientific insights. The unforeseen differences may in time be overcome by incremental alignment; however, great care must be taken to avoid unnecessary disruption to working practices and functioning systems.

3 Requirement Analysis

We can now consider the requirements of RIs on a per-topic basis.

3.1 Identification and Citation

Identification of data (and associated metadata) throughout all stages of processing is central in any RI and can be ensured by allocating unique and persistent digital identifiers

(PIDs) to data objects throughout the data lifecycle. PIDs allow unambiguous references to be made to data during curation and cataloguing. They are also a necessary requirement for correct citation (and hence attribution) of data by end-users. Environmental RIs are often built on a large number of distributed observational or experimental sites, run by hundreds of scientists and technicians, financially supported and administered by a large number of institutions. If this data is shared under an open access policy, it becomes therefore very important to acknowledge the data sources and their providers.

The survey of ENVRI RIs found great diversity between RIs regarding their practices. Most apply file-based storage for their data, rather than database technologies, which suggests that it should be relatively straightforward to assign PIDs to a majority of the RI data objects. A profound gap in knowledge about what persistent and unique identifiers are, what they can be used for, and best practices regarding their use, emerged, however. Most identifier systems used are based on handles (DOIs from DataCite[3] most common, followed by ePIC PIDs[4]), but some RIs rely on formalised file names. While a majority see a strong need for assigning PIDs to their "finalised" data (individual files and/or databases), few apply this to raw data, and even fewer to intermediate data—indicating PIDs are not used in workflow administration. Also, metadata objects are seldom assigned PIDs.

Currently, researchers refer to datasets in publications using DOIs if available, and otherwise provide information about producer, year, report and number either in the article text or in the bibliography. A majority of RIs feel it is absolutely necessary to allow unambiguous references to be made to specified subsets of datasets, preferably in the citation, while few find the ability to create and later cite collections of individual datasets as important. Ensuring that credit for producing (and to a lesser extent curating) scientific datasets is "properly assigned" is a common theme for all RIs—not least because funding agencies and other stakeholders require such performance indicators, but also because individual PIs want and need recognition of their work. Connected to this, most RIs have strategies for collecting usage statistics for their data products, i.e., through bibliometric searches (quasi-automated or manual) from scientific literature, but thus often rely on publishers indexing also data object DOIs.

The use of persistent and unique identifiers for both data and metadata objects throughout the entire data lifecycle needs to be encouraged, e.g. by providing training and best-use cases. There is strong support for promoting "credit" to data collectors, through standards of data citation supporting adding specific sub-setting information to a basic (e.g. DOI-based) reference. Demonstrating that this can be done easily and effectively, and that data providers can trust that such citations will be made, will be a priority, as it will lead to adoption and improvement of citation practices.

A key issue is adoption of appropriate steps in working practices. Where these are exploratory or innovative the citation of underpinning data may be crucial to others verifying the validity of the approach and to later packaging for repeated application. Once a working practice is established, it should be formalised, e.g. as a workflow, and packaged, e.g. through good user interfaces, so that as much of the underpinning record keeping e.g. citation, cataloguing and provenance is automated. This has two positive

[3] https://datacite.org/.

[4] https://www.pidconsortium.eu.

effects: it enables practitioners to focus on domain-specific issues without distracting record keeping chores, and it promotes a consistent solution to be incrementally refined. For these things to happen there have to be good technologies, services and tools supporting each part of these processes, e.g. data citations being automatically and correctly generated as suggested by Buneman et al. [7]. Similarly, constructing immediate payoffs for practitioners using citation, as suggested by Myers et al. [8], will increase the chances of researchers engaging with identification at an earlier stage.

Many researchers today access and therefore consider citing individual files. This poses problems if the identified files may be changed, the issue of fixity. Many research results and outputs depend on very large numbers of files and simply enumerating them does not yield a comprehensible citation. Many derivatives depend on (computationally) selected parts of the input file(s). Many accesses to data are via time varying collections, e.g. catalogues or services, that may yield different results or contents on different occasions—generically referred to as databases. Some results will deal with continuous streaming data. Often citations should couple together the data sources, the queries that selected the data, the times at which those queries were applied, the workflows that processed these inputs and parameters or steering actions provided by the users (often during the application of the scientific method) that potentially influenced the result. All of these pose more sophisticated demands on ata identification and citation systems. At present they should at least be considered during the awareness raising proposed above. In due course, those advanced aspects that would prove useful to one or more of the RI communities should be further analysed and supported.

3.2 Curation

When the RIs in the ENVRI community were surveyed during the ENVRIplus project (in 2016), many did not have fully detailed plans for how they would curate their data and did not yet have complete Data Management Plans; in the years since, many of these RIs have made significant progress as they move into the implementation phases of their respective developments.

Only one RI mentioned OAIS[5] (the ISO/IEC 14721 standard for curation); this may be because it is not much used, and when it is the implementations are very varied since it is really an overview architecture rather than a metadata standard. With regard to the metadata standards used or required by the RIs, several used ISO19115/INSPIRE, one used CERIF, and one used Dublin Core; of these standards, only CERIF explicitly provides curation information. None mentioned metadata covering software and its curation except EPOS (using CERIF). A few use Git[6] to manage software. Most have no curation of software nor plans for this.

Possibly due to the early stage of some RIs, the requirements for curation were not made explicit, for example, none of the RIs (who responded) had appropriate metadata and processes for curation. It is known that EPOS has plans in place and there are indications of such planning for some of the others. Since curation often underpins

[5] http://www.dcc.ac.uk/resources/curation-reference-manual/chapters-production/using-oais-reference-model-curation.

[6] https://git-scm.com/.

validation of the quality of scientific decisions and since environmental sciences observe phenomena that do not repeat in exactly the same form, the profile of curation needs raising.

Curation requirements validate the need to develop curation solutions but do not converge on particular technical requirements. Some further issues arise. These are enumerated below:

- The need for intellectual as well as ICT interworking between these closely related topics: Identification and Citation, Curation, Cataloguing and Provenance is already recognised. Their integration will need to be well supported by tools, services and processing workflows, used to accomplish the scientific methods and the Curation procedures. The need for this combination for reproducibility is identified by Belhajjame et al. with implementations automatically capturing the context and synthesising virtual environments [9].
- As above, it is vital to support the day-to-day working practices and innovation steps that occur in the context of Curation with appropriate automation and tools. This is critical both to make good use of the time and effort of those performing Curation, and to support innovators introducing new scientific methods with consequential Curation needs.
- Curation needs to address preservation and sustainability; carefully preserving key information to underwrite the quality and reproducibility of science requires that the information remains accessible for a sufficient time. This is not just the technical challenge of ensuring that the bits remain stored, interpretable and accessible. It is also the socio-political challenge of ensuring longevity of the information as communities' and funders' priorities vary. This is a significant step beyond archiving, which is addressed in EUDAT for example with the B2SAFE service.

One aspect of the approach to sustainable archiving is to form federations with others undertaking data curation, as suggested by OAIS. Federation arrangements are also usually necessary in order that the many curated sources of data environmental scientists need to use are made conveniently accessible. Such data-intensive federations (DIF) underpin many forms of multi-disciplinary collaboration and supporting them well is a key step in achieving success. As each independently run data source may have its own priorities and usage policies, often imposed and modified by its funders, it is essential to set up and sustain an appropriate DIF for each community of users. Many of the RIs deliver such federations, today without a common framework to help them, and many of the ENVRIplus partners are members of multiple federations.

3.3 Cataloguing

Regarding the possible items to be managed in catalogues, the RIs surveyed showed interest in:

- Observation systems and lab equipment: most RIs manage equipment which requires management (e.g. scheduling, maintenance and monitoring) and some of them are

managing or would like to manage this with an information system. Some are already using a standardised approach (OGC/SWE and SSN).
- Data processing procedures and systems, software: a very few or none mentioned an interest in supporting this in a catalogue. We observe, however, that this may be necessary as part of the provision for provenance and as an aid for those developing or formalising new methods.
- Observation events: not explicitly mentioned as a requirement most of the time. Again, this need may emerge when provenance is considered.
- Physical samples: mentioned by a few especially in the biodiversity field.
- Processing activities: not explicitly mentioned.
- Data products or results: widely mentioned as being done by existing systems (EBAS, EARLINET, CLOUDNET, CKAN, MAdrigal and DEIMS) and widely standardised (ISO/IEC 191XX). Compliance is sometimes required with the INSPIRE directive; support for this in the shared common subsystems would prove beneficial.
- Publications: widely mentioned. However, very few manage the publications on their own. Links for provenance between publications and datasets are quite commonly required.
- Persons and organisations: not explicitly mentioned. However, this is reference information, which is required for the other described items (e.g. datasets and observation systems) and for provenance (contact points).
- Research objects or features of interest: mentioned once as feature of interest (airports for IAGOS).

As a consequence, the following three categories of catalogues are cited in the requirements collection:

- Reference catalogues, which are not developed by ENVRIplus or within RIs but are pre-existing infrastructures containing reference information to be used. They can also be considered as gazetteer, thesaurus or directories. Among them we consider catalogues for people and organisations, publications, research objects, and features of interest.
- Federated catalogues, which are pre-existing and partly harmonised in an RI but could be federated by ENVRIplus. Among them we consider data products, results, observation systems and lab equipment, physical samples, data processing procedures and systems, and software components metadata.
- Finally, activity records, observation events, processing activities, and usage logs can be considered.

There are a wide variety of items that could be catalogued, from instruments and deployments at the data acquisition stage, right through every step of data processing and handling, including the people and systems responsible, up to the final data products and publications made available for others to use [10]. Most responding RIs pick a small subset of interest, but it is possible that a whole network of artefacts needs cataloguing to facilitate Provenance, and many of these would greatly help external and new users find and understand the research material they need. There is a similar variation in the kinds of information, metadata, provided about catalogue entries. Only EPOS has a

systematic approach in its use of CERIF, though many have commonalities developing out of the INSPIRE directive (EU Directive 2007/2/EC)[7]. So again we will consider a few implications:

- A critical factor that emerged in general requirements discussions was the need to easily access data. This clearly depends on good query systems that search the relevant catalogues and couple well with data handling and provenance recording. The query system is closely coupled with catalogue design and provision, but it also needs integration with other parts of the system. Euro-Argo identified a particular version of data access—being able to specify a requirement for a repeating data feed.
- Catalogues are a key element in providing convenient use of federations of resources. It is probably necessary to have a high-level catalogue that identifies members of the federation and the forms of interaction, preferably machine-to-machine, they support. Initially users may navigate this maze and handle each federation partner differently, but providing a coherent view and a single point of contact has huge productivity gains. It is a moot point whether this requires an integrated catalogue or query systems that delegate sub-queries appropriately. This is another example where effective automation can greatly improve the productivity of all the RI's practitioners; those that support the systems internally and maintain quality services, and those who use the products for research and decision making. It is anticipated that federations will grow incrementally and that the automation will advance to meet their growing complexity and to deliver a holistic and coherent research environment where the users enjoy enhanced productivity. This will depend on catalogues holding the information needed for that automation as well as the information needed for RI management and end-user research.
- Once again there may be some merit in making the advantages of catalogues evident in the short-term, e.g. by coupling catalogue use with operations that user want to perform, such as: having selected data via a catalogue, moving it or applying a method to each referenced item. Similarly, allowing the users some free-form additions and annotations to catalogue entries that help them pursue their own goals may be helpful.

3.4 Processing

Data processing (or analytics) is an extensive domain, including any activity or process that performs a series of actions on dataset(s) to distil information [11]. It may be applicable at any stage in the data lifecycle from quality assurance and event recognition close to data acquisition to transformations and visualisations to suit decision makers as results are presented. Data analytics methods draw on multiple disciplines, including statistics, quantitative analysis, data mining, and machine learning. Very often these methods require compute-intensive infrastructures to produce their results in a suitable time, because of the data to be processed (e.g. huge in volume or heterogeneity) and/or because of the complexity of the algorithm/model to be elaborated/projected. Moreover, these methods being devised to analyse datasets and produce other "data"/information (than can be considered a dataset) are strongly characterised by the "typologies" of their

[7] https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32007L0002.

inputs and outputs. In some data-intensive cases, the data handling (access, transport, IO and preparation) can be a critical factor in achieving results within acceptable costs.

As largely expected, RIs' needs with respect to datasets to be processed are quite diverse because of the diversity in the datasets that they deal with. Datasets and related practices are diverse both across RIs and within the same RI. For instance, in EPOS there are many communities each having its specific typologies of data and methodologies (e.g. FTP) and formats (e.g. NetCDF) for making them available. Time series and tabular data are two very commonly reported types of dataset to be processed yet they are quite abstract. In what concerns "volume", datasets vary from a few kilobytes to terabytes and beyond. In the large majority of cases datasets are made available as files while few infrastructures have plans to make or are making their data available through OGC services, e.g. ACTRIS. The need to homogenise and promote state-of-the-art practices for data description, discovery and access is of paramount importance to providing RIs with a data processing environment that makes it possible to easily analyse dataset(s) across the boundaries of RI domains.

Considering actual processing itself, it emerged that RIs are at diverse levels of development and that there is a large heterogeneity. For instance, the programming languages currently in use by the RIs range from Python, Matlab and R to C, C++, Java, and Fortran. The processing platforms available to RIs range from a few Linux servers to the HPC approaches exploited in EPOS. Software in use or produced tends to be open source and freely available. In the majority of cases there is almost no shared or organised approach to make available the data processing tools systematically both within the RI and outside the RI. One possibility suggested by some RIs is to rely on OGC/WPS for publishing data processing facilities.

Any common processing platform should be open and flexible enough to allow: (a) scientists to easily plug-in and experiment with their algorithms and methods without bothering with the computing platform; (b) service managers to configure the platform to exploit diverse computing infrastructures; (c) third-party service providers to programmatically invoke the analytics methods; and (d) engineers to support scientists executing existing analytic tasks eventually customising/tuning some parameters without requiring them to install any technology or software.

Regarding the output of processing tasks, we can observe that the same variety characterising inputs as being there for outputs also. In this case, however, it is less well understood that there is a need to make these data available in a systematic way, including information on the entire process leading to the resulting data. In the case of EMBRC it was reported that the results of a processing task are to be made available via a paper while for EPOS it was reported that the dataset(s) are to be published via a shared catalogue describing them by relying on the CERIF metadata format.

In many cases, but by no means all, output resulting from a data processing task should be 'published' to be compliant with Open Science practices. A data processing platform capable of satisfying the needs of scientists involved in RIs should offer an easy to use approach for having access to the datasets that result from a data processing task together. As far as possible it should automatically supply the entire set of metadata characterising the task, e.g. through the provenance framework. This would enable scientists to properly interpret the results and reduce the effort needed to prepare for curation. In

cases where aspects of the information are sensitive, could jeopardise privacy, or have applications that require a period of confidentiality, the appropriate protection should be provided.

Only a minority of the RIs within ENVRIplus contributed information about statistical processing during requirements gathering. Unsurprisingly given the diversity of the component RIs, there were a variety of different attitudes to the statistical aspects of data collection and analysis. One RI (IS-ENES-2) felt that data analysis (as opposed to collection) was not their primary mission, whereas for others (e.g. within EMBRC) reaching conclusions from data is very much their primary purpose.

As environmental data collection is the primary aim of many of the RIs it appears that day-to-day consideration of potential hypotheses underlying data collection is not undertaken. Hypothesis generation and testing is for scientific users of the data and could take many forms. However, some RIs (e.g. LTER and ICOS) stressed that general hypotheses were considered when the data collection programmes and instruments were being designed especially if the data fed into specific projects. Hypotheses could be generated after the fact by users after data collection and indeed this would be the norm if data collection is a primary service to the wider scientific community.

RIs can be collecting multiple streams of data often as time series, thus there is the potential to undertake multivariate analysis of the data. Again unsurprisingly given the diversity of science missions, there was no consistency in approaches. Data could be continuous and discrete, be bounded by its very nature or have bounds enforced after collection. Datasets are potentially very voluminous; total datasets with billions of sample points might be generated. Most analysers will be engaging in formal testing of hypotheses rather than data mining, although the latter was not necessarily ruled out. Many RIs had or are going to implement outlier or anomaly detection on their data.

The wide scope of potential contexts in which processing could be applied: from quality assurance close to data acquisition to transformations for result presentation (and every research, data-management or curation step in between) makes this a complex factor to consider. User engagement with this topic also varies validly between two extremes: those who use a pre-packaged algorithm in a service almost unknowingly as part of a well-formalised, encapsulated, established method they use, to those who are engaged in creating and evaluating new algorithms for innovative ways of combining and interpreting data. Clearly, both continua are valid and any point in each continuum needs the best achievable support for the context and viewpoint. With such diversity it is clear that a one-size-fits-all approach is infeasible. This conclusion is further reinforced by the need to exploit the appropriate computational platforms (hardware architectures, middleware frameworks and provision business models) to match the properties of the computation, and the priorities of the users given their available resources. If such matching is not considered it is unlikely that all of the developing research practices will be sustainable in an affordable way. For example, too much energy may be used or the call on expert help to map to new platforms may prove unaffordable. Such issues hardly rise to the fore in the early stages of an RI or a project. So again, we note forces that will cause the understanding and nature of requirements to evolve with time. This leads to the following follow-up observations:

- The packaging of computations and the progressive refinement of scientific methods are key to productivity and to the quality of scientific conclusions. Consequently, as far as possible processing should be defined and accessed by high-level mechanisms. This allows a focus on the scientific domain issues and it leaves freedom for optimised mappings to multiple computational platforms. This protects scientific intellectual investment, as it then remains applicable as the computational platforms change. This will happen as their nature is driven by the much larger entertainment, media, leisure and business sectors. The higher-level models and notations for describing and organising processing also facilitates optimisation and automation of chores that otherwise will distract researchers and their supporters.
- Providing support for innovation in this context is critical. Without innovation the science will not advance and will not successfully address today's societal challenges. It requires support for software development, testing, refinement, validation and deployment conducted by multi-site teams engaging a wide variety of viewpoints, skills and knowledge. For the complex data-intensive federations the environmental and Earth sciences are dealing with, this involves new intellectual and technological territory. Alliances involving multiple RIs and external cognate groups such as EUDAT, PRACE and EGI, may be the best way of gathering sufficient resources and building the required momentum.

3.5 Provenance

For modern data-driven science there is a pressing need to capture and exploit good provenance data. Provenance, the records about how data was collected, derived or generated, is crucial for validating and improving scientific methods and is a key aspect of making data FAIR. It enables convenient and accurate replay or re-investigation and provides the necessary underpinning when results are presented for judging the extent to which they should influence decisions whether for hazard mitigation or paper publication. It provides a foundation for many activities of import to RIs, such as attributing credit to individuals and organisations, providing input to diagnostic investigations, providing records to assist with management and optimisation and preparing for curation. All RIs will need to perform these functions and consequently the e-infrastructures they depend on will need to support provenance collection and use as well.

Most RIs already consider provenance data as essential and are interested in using a provenance recording system. Among all of the nine RIs who gave feedback about provenance only two already had a data provenance recording system embedded in their data processing workflows. EPOS uses the dispel4py workflow engine in VERCE[8], which is based on and is able to export to PROV-O whereas in future it is planned to use the CERIF data model and ontology instead [12]. IS-ENES2 instead does not specify which software solution is applied but mentions: the use of community tools to manage what has been collected from where, and what is the overall transfer status to generate provenance log files in workflows [13]. Some, such as SeaDataNet and Euro-Argo, interpret provenance as information gathered via metadata about the lineage

[8] http://www.verce.eu/AboutVerce/RelatedProjects.php.

data with tools like Geonetwork based on metadata standards like ISO19139[9], but the information gathered is not sufficient to reproduce the data as the steps of processing are not documented in enough detail. Other RIs, such as ICOS and LTER, are already providing some provenance information about observation and measurement methods used within the metadata files but are aware that a real tracking tool still needs to be implemented. IAGOS is using the versioning system GIT for code but not the data itself. A versioning system can only be seen as a part of the provenance information sought.

On which information is considered to be important, the answers range from versioning of data to the generation of data and modification of the data as well as on who, how and why data is used. So there seems to be two interpretations about what provenance should comprise: should it enable the community to follow the data 'back in time' and see all the steps that happened from raw data collection, via quality control and aggregation to a useful product, or should it enable the data provider as a means of tracking the usage of the data, including information about users in order to understand the relevance of the data and how to improve their services? These two roles for metadata may be served by the same provenance collecting system. The provenance data is then interpreted via different tools or services.

Regarding the controlled vocabulary used for the descriptions of the steps for data provenance, some RIs already use research specific reference tables and thesauri like EnvThes and SeaDataNet common vocabularies. There is great interest among the RIs to get clear recommendations from ENVRIplus about the information range provenance should provide. This includes drawing an explicit line between metadata describing the 'dataset' and provenance information. Also it should be defined clearly whether usage tracking should be part of provenance. It is considered as being very important to get support on automated tracking solutions and or provenance management APIs to be applied in the specific e-science environments. Although there are some thesauri already in use there is a demand for getting a good overview of the existing vocabularies and ontologies that are ready to use or that need to be slightly adapted for specific purposes.

At present, the need for and benefits of provenance provision are only recognised by some RIs. In abstract, we are sure that most scientists appreciate the value of provenance, but they tend to think of it as a painful chore they have to complete when they submit their final, selectively chosen data to curation. They often only do this when their funders or publishers demand it. That culture is inappropriate. For many RIs they are in the business of collecting and curating primary data and commonly required derivatives. Clearly, they want to accurately record the provenance of those data, as a foundation for subsequent use and to achieve accountable credit. For environmental and Earth scientists use of provenance throughout a research programme can have significant benefits. During method development it provides ready access to key diagnostic and performance data, and greatly reduces the effort required to organise exactly repeated re-runs; a frequent chore during development. As they move to method validation they have the key evidence to hand for others to review. When they declare a success and move to production, the provenance data informs the systems engineers about what is required and can be exploited by the optimisation system. Once results are

[9] https://www.iso.org/standard/32557.html.

produced using the new method these development-time provenance records underpin the provenance information collected during the production campaign.

The RIs survey reported very different stages of adoption, and when there was adoption it did not use the same solutions or standards—this was almost always related to data acquisition rather than the use of data for research. The change in culture among researchers may be brought about by ENVRIplus through a programme of awareness raising and a well-integrated compendium of tools. The latter may be more feasible if the development of the active provenance framework is amortised over a consortium of RIs. This leads to similar observations to those given above:

3.6 Optimisation

Environmental and Earth sciences now rely on the acquisition of great quantities of data from a range of sources for building the complex models. The data might be consolidated into a few very large datasets, or dispersed across many smaller datasets; it may be ingested in batch or accumulated over a prolonged period. Although efforts are underway to store data in common data stores, to use this wealth of data fast and effectively, it is important that the data is both optimally distributed across a research infrastructure's data stores, and carefully characterised to permit easy retrieval based on a range of parameters. It is also important that experiments conducted on the data can be easily compartmentalised so that individual processing tasks can be parallelised and executed close to the data itself, so as to optimise the use of resources and provide swift results for investigators.

Perhaps more so than the other topics, optimisation requirements are driven by the specific requirements of those other topics, e.g. time-critical requirements for data processing and management [14]. For each part of an infrastructure in need for improvement, we must consider what it means for the part to be optimal and how to measure that optimality.

In the context of common services for RIs, it is necessary to focus on certain practical and broadly universal technical concerns, generally those being to do with the movement and processing of data. This requires a general understanding of what bottlenecks exist in the functionality of (for example) storage and data management subsystems, understanding of peak volumes for data access, storage and delivery in different parts of the infrastructure, understanding of computational complexity of different data-processing workflows, and understanding of the quality of service requirements researchers have for data handling in general [15]. Many optimisation problems can be reduced down to ones of data placement, in particular of data staging, whereby data is placed and prepared for processing on some computational service (whether that is provided on a researcher's desktop, within an HPC cluster or on a web server), which in turn concerns the further question of whether data should be brought to where they can be best computed, or instead computing tasks be brought to where the data currently reside. Given the large size of many RI's primary datasets, bringing computation to data is appealing, but the complexity of various analyses also often requires supercomputing-level resources, which require the data be staged at a computing facility such as are brokered in Europe by consortia such as PRACE. Data placement is reliant however on data accessibility, which is not simply based on the existence of data in an accessible location, but is also

based on the metadata associated with the core data that allows it to be correctly interpreted; it is based on the availability of services that understand that metadata and can so interact (and transport) the data with a minimum of manual configuration or direction.

Experience shows that as data-handling organisations transition from pioneering to operations, many different reasons for worrying about optimisation emerge. These are addressed by a wide variety of techniques, so that investment in optimisation is usually best left until the RI or RI community can establish what they want to be optimised and the trade-offs that they would deem acceptable. Very often there are significantly different answers from different members of a community. The RI's management may need to decide on compromises and priorities.

Optimisation needs to look beyond individuals and single organisations. When looking at overall costs or energy consumption in a group of RIs or the e Infrastructures they use, tactics may consider the behaviour of a data-intensive federation. For example, when data is used from remote sites, or is prepared for a particular class of uses, the use of caching may save transport and re-preparation costs, and accelerate the delivery of results. However, the original provider organisation needs to have accountable evidence that their data is being used indirectly, and the caching organisation needs its compute and storage costs amortised over the wider community.

3.7 Community Support

We define community support as part of the RI concerned with managing, controlling and tracking users' activities within an RI and with supporting all users to conduct their roles in their communities. It includes many miscellaneous aspects of RI operations, including for example authentication, authorisation and accounting, the use of virtual organisations, training, and help-desk activities.

The questions asked of RIs focused on 3 aspects: a) functional requirements; b) non-functional requirements (e.g. privacy, licensing and performance); and c) training.

The following is a summary of the main functional requirements expressed by the RIs (not all apply to all RIs):

- Data Portal: a data portal was frequently requested by RIs. Many RIs already have their own data portal, and some of the others are in the process of developing one. Data portals provide (a single point of) access to the system and data both for humans and machines (via APIs). Commonly requested functionalities include access control (for example, IS-ENES2 currently uses OAuth2, OpenID, SAML and X.509 for AAI management) and discovery of services and data facilities (e.g. metadata-based discovery mechanisms).
- Accounting: the tracking of user activities, which is useful for analysing the impact of the RI, is commonly requested. For example, EMBRC records where users are going, what facilities they are using, and the number of requests. The EMBRC head office will in the future provide a system to analyse resource DOIs, metrics for the number of yearly publications and impact factor, and questionnaires submitted by users about their experience with their services. LTER plans to track the provenance of the data, as well as its usage (e.g. download or access to data and data services). DEIMS, for example is planning that statistics about users will be implemented, mainly to allow

for a better planning of provided services. Features will be implemented by exploiting EUDAT services, e.g. provenance support of B2SHARE to track data usage. Google analytics is currently used to track the usage of the DEIMS interface.

- Issue tracker: ACTRIS has recently introduced an issue tracker to link data users and providers, and to follow up on feedback on datasets in response to individual requests.
- Community software: EPOS is in the process of deciding which private software to use and how to integrate it in the data portal. In LTER, the R statistical software and different models (e.g. VSD+ dynamic soil model, LandscapeDNDC regional scale process model for simulating biosphere-atmosphere-hydrosphere exchanges.) are provided.
- Wiki: a wiki is often used to organise community information, and as a blackboard for collaborative work for community members (e.g. to add names and responsibilities to a list of tasks to be done). Sometimes, it is also used to keep track of the progress on a task, both for strategic and IT purposes. FAQ pages (and other material targeting a more general audience, or outreach materials for educational institutes) are a special type of wiki page describing more technical aspects of data handling and data products, and also a system for collecting user feedback.
- Mailing lists, twitter & Forums are intended to facilitate communication to and from groups of community members. Forums and mailing lists can be interlinked so that any message in the mailing list is redirected to the forum and vice-versa.
- Files and image repositories represent shared spaces where members and stakeholders can upload/download and exchange files. They are also a fundamental tool for storing and categorising images and other outreach materials.
- Shared calendars keep track and disseminate relevant events for community members.
- Tools to organise meetings, events and conferences should handle all the aspects of a conference/meeting: programme, user registration, deadlines, document submission and dissemination of relevant material. Tools like Indico are currently popular.
- Website: The purpose of the website is to disseminate community relevant information to all stakeholders. The website should not contain reserved material but only publicly accessible material (e.g. documents and presentations for external or internal stakeholders, images for press review). The website should also include news and interactions from social networks. The website should be simple enough to allow almost anyone with basic IT skills to add pages, articles, images. A simple CMS (content management system) is the most reasonable solution (e.g. Wordpress, Joomla).
- Teleconferencing tools: Communication with all stakeholders (internal and external) is also carried on through teleconferencing. For this purpose, good quality tools (e.g. screen sharing, multi-user, document exchange, and private chat) are needed. Popular tools include Adobe Connect, Web Ex, GoToMeeting, Google Hangout and Skype.
- Helpdesk & Technical support: For example, the data products that ICOS produces are complex and often require experience of, and detailed knowledge about the underlying methods and science to be used in an optimal way. Technical support must be available to solve any problem. The ICOS Thematic Centres (for Atmosphere, Ecosystems and Ocean) are ready to provide information and guidance for data users. If needed, requests for information may also be forwarded to the individual observation stations. The mission of ICOS also comprises a responsibility to support producers of derived

products (typically research groups performing advanced modelling of greenhouse gas budgets) by providing custom-formatted "data packages".

The non-functional requirements of the RIs that were most frequently referred to were:

- Performance: RIs need robust, fast-reacting systems, which offer security and privacy. Moreover, they need good performance for high data volumes.
- Data policy and licensing constraints: The data produced by some communities has licensing constraints that restrict access to a certain group of users. For example, while ICOS will not require its users to register in order to use the data portal or to access and download data, it plans to offer an enhanced usage experience to registered users. This will include automatic notifications of updates of already downloaded datasets, access to additional tools, and the possibility to save personalised searches and favourites in a workspace associated with a user's profile. Everyone who wishes to download ICOS data products must also acknowledge the ICOS data policy and data licensing agreement (registered users may do so once, while others must repeat this step every time).

Training activities within ENVRIplus communities can be categorised as follows:

- No training plan: The majority of ENVRIplus RI communities do not have a common training plan at the moment.
- No community-wide training activities: For example:
  - Within SIOS, many organisations have their own training activities. Training is provided to students or scientists. For example, The University Centre in Svalbard (UNIS) has its own high-quality-training programme on Arctic field security (i.e., how to operate safely in an extreme cold climate and in accordance with environmental regulations) for students and scientists.
  - Within ACTRIS, each community has its own set of customised training plans. Courses and documentation are made available online, for example for training on how to use the data products. Their preferred methods for delivering training are through the community website or through targeted sessions during community specific workshops. ACTRIS also considers organising webinars.
  - ICOS does not have a common training plan at the moment. The Carbon Portal organises occasional training events, e.g. on Alfresco DMS (Document Management System used by ICOS RI). The different Thematic Centres periodically organise training for their respective staff and in some cases also for data providers (station PIs). ICOS also (co-)organises and/or participates in summer schools and workshops aimed at graduate students and postdocs in the relevant fields of greenhouse gas observational techniques and data evaluation. Representatives of ICOS have participated in training events organised by EUDAT, e.g. on PID usage and data storage technology. The method of delivering training through one- or two-day face-to-face workshops concentrated on a given topic and with a focus on hands-on activities is probably the most effective. This should also be backed up by webinars (including recordings from the workshops) and written materials.

- A community training plan is under development: A number of communities are in the process of developing a community training plan. For example:
  - LTER plans the development of a community-training plan. Within LTER Europe, the Expert Panel on Information Management is used to exchange information on a personal level and to guide developments such as DEIMS to cater for user needs. LTER Europe also provides dissemination and training activities to selected user groups. Training activities will enhance the quality of the data provided, by applying standardised data quality control procedures for defined data sets.
  - For EPOS, training is part of its communication plan.

- An advanced system is in place for training activities:
  - Within IS-ENES2, workshops are organised from time to time. Also, communities communicate about the availability of training courses and workshops organised by HPC centres (PRACE) or EGI.
  - Within EMBRC, a Training web portal is provided, offering support to training organisers to advertise and organise courses.

The above list covers virtually all of the facilities for communication, information sharing, organisation and policy implementation that a distributed community of collaborating researchers and their support teams might expect—and they normally expect those facilities to be well integrated and easily accessed wherever they are from a wide range of devices. However, care should be taken to consider the full spectrum of end users. A few may be at the forefront of technological innovations but the majority may be using very traditional methods, because they work for them. Investment is only worthwhile if it is adopted and benefits the greater majority of such communities, taking into account their actual preferences.

There may be two key elements missing in the context of ENVRIplus, which focuses on achieving the best handling and use of environmental data:

- Workspaces that can be accessed from anywhere and are automatically managed, in which individuals or groups can store and organise the data concerned with their work in progress: e.g. test data sets, sample result sets, intermediate data sets, results pending validation, results pending publication. Since environmental researchers have to work in different places, such as in field sites, in different laboratories and institutions, they need to control these logical spaces, which may be distributed for optimisation or reliability reasons. These are predominantly used to support routine work but can also be used for innovation. This includes intelligent sensors requiring access to a variety of logical spaces for their operations.
- Development environments that can be accessed from most workstations and laptops, and that facilitates collaborative innovation and refinement of the scientific methods and data handling. Sharing among a distributed community, testing, management of versions and releases and deployment aids would be expected.

To a lesser or greater extent every RI will depend on a mix of roles and viewpoints. Community support needs to recognise and engage with these multiple viewpoints as

well as help them to work together. This is particularly challenging in the distributed environments and federated organisations underpinning many RIs. At least training and help desk organisation will need to take these factors into account. Productivity will come from each category being well supported. Significant breakthroughs will depend on the pooling of ideas and effort across category boundaries.

3.8 Cross-Cutting Requirements

There are a few additional requirements that appear in the analysis of RI needs that have aspects of improving usability to improve the experience and productivity of users and the teams who support them. In part, they are better packaging of existing or planned facilities and in part they are intended to deliver immediate benefits to keep communities engaged and thereby, improve take up and adoption of RI products.

- Boundary crossing. The participating communities experience boundaries between the different roles identified above (see Sect. 3.7), between disciplines, sub-disciplines and application domains, and between organisations. This can be stimulated by:
  - Organising ad hoc think tanks so that it brings together (virtually) participants from across the boundaries and stimulates them to think and work together on relevant topics, e.g. by bringing in suitable experts and setting up suitable practical challenges to be addressed during the course. This requires elapsed time, and allocation of both training effort and trainee time, so the target understanding that the course will deliver has to be carefully chosen.
  - Establishing suitable agile development processes where people work intensely together on a common issue with a carefully set goal. Then assimilating the results and building on the networks provided.
  - Delivering services and tools well suited to each role and organisational context.
  - Arranging workspaces that facilitate such collaborative behaviour while ideas are being developed and formulated. This requires those involved to have control over the release and sharing of the material they work on. Individuals may be involved in several groups, probably with different roles.

- Integrated communication facilities. The individual elements of communication for distributed participants in an RI need to be conveniently integrated. There are several potential solutions in this area. It may help if at least one well-integrated one were run to be available for RIs, project participants and ENVRIplus. This needs to present views that work well for each category of practitioner. Some of the selected use cases in Theme 2 may serve to achieve this.
- Exemplars and early benefits. The development of exemplars of effective methods and software or services that support them is key to spreading ideas, testing them in new contexts and developing buy in. This will be helpful in the training and outreach programme. It is also vital as part of the process of delivering as early as possible benefits to the active researchers and other practitioners. If we can deliver immediate benefits they will not have to struggle for so long investing unproductive time in tedious workarounds.

- Data access interfaces. Researcher and others managing data-driven processes spend a great deal of time, identifying data they want, arranging to be permitted access, arranging transfers, arranging local storage, arranging onward shipment to computation resources if necessary and returning storage resources when they have finished. If this is packaged as a convenient operation their work is simplified and more productive. The parts of such a process are all being built, but delivering an integrated solution that just works would be a large benefit. It needs the provision of a user's or group's workspace. It needs a means of identifying the required data. Once deployed, it can be grown in small increments, taking the users along an improving path. They might prioritise some of the following:

  – Identification using queries over associated metadata (in the identity registries or in catalogues (see Sect. 3.2)).
  – Extension of the operations that are easily applied to the accessed data (we have found visualisation particularly relevant).
  – Handling batches of data consistently at the same time (the tea tray metaphor).
  – Handling intermediate (transient) results with various aids for handling them in bulk and for clearing up afterwards.
  – Promoting selected results to properly identified and citable.
  – Arranging for their data to be published or curated.

4 Conclusion

A general conclusion that can be drawn from the information acquired from the RIs is that there are more differences than commonalities between the RIs—the RIs are all at different stages in their development and have different organisational status, from well-established and operational to RIs still in their definition phase. Moreover, some are heavily distributed with heterogeneous networks of sensors and network services in different countries making different kinds of measurement or observation, while other RIs have one single observing platform and one central hub for data. Nevertheless, it is still possible to identify a number of key common concerns. These include:

- The need to achieve data harmonisation, i.e., consistency of representation, interpretation and access, both within and between RIs.
- The need for RIs to learn from one another and pool efforts in order to accelerate and harmonise delivery of data services and working practices that efficiently support each stage of the scientific data lifecycle, from data acquisition to delivery of actionable derived information.
- Help with facing the challenge of sustainably delivering data services immediately to meet current RI priorities while considering longer-term issues and technology trends.

The ability to describe different processes from multiple viewpoints in a standard way helps facilitate the collaboration between RIs and alignment of their activities. The ENVRI Reference Model (RM) [16] provides a conceptual model to this, enabling RI communities to discuss and see where improvements in data processing in the RIs are

possible and required. The RM is a living model that has been developed on the basis of evaluating RIs within the ENVRI community. Data science solutions that can fulfil the identified requirements can be expressed in terms of the RM and then projected onto RIs in order to help in optimising their data lifecycles.

Atkinson et al. [5] provide an in-depth analysis of the state of the RIs and the technologies they used as of mid 2016, providing a number of recommendations. Atkinson et al. stress that the diversity within and between RIs and the complexity of the RIs, involving many different roles, require effective communication and collaboration to address. Data sharing and governance of the data is essential to RI operation and to the production of valuable science, and needs to be considered with ample allocation of resources and attention as a main priority when setting up and governing RIs. This requires training of staff and education of future scientists. Shared developments in sustainable software and platforms for performing data-driven (environmental) science are also needed to minimise costs and increase the sustainability of the RIs.

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. Hey, T., Tansley, S., Tolle, K. (eds.): The fourth paradigm: data-intensive scientific discovery. Microsoft Research (2009)
2. Zhao, Z., et al.: Reference model guided system design and implementation for interoperable environmental research infrastructures. In: 2015 IEEE 11th International Conference on e-Science, Munich, Germany, pp. 551–556. IEEE (2015). https://doi.org/10.1109/eScience.2015.41
3. Wilkinson, M., Dumontier, M., Aalbersberg, I., et al.: The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data **3**, 160018 (2016). https://doi.org/10.1038/sdata.2016.18
4. 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), San Diego, CA, USA, pp. 277–280. IEEE (2019). https://doi.org/10.1109/escience.2019.00038. https://zenodo.org/record/3462816
5. Atkinson, M., et al.: D5.1 A consistent characterisation of existing and planned RIs. H2020 ENVRIplus Project (2016). http://www.envriplus.eu/wp-content/uploads/2016/06/A-consistent-characterisation-of-RIs.pdf
6. Tanhua, T., et al.: Ocean FAIR data services. Front. Mar. Sci. **6**, 440 (2019). https://doi.org/10.3389/fmars.2019.00440
7. Buneman, P., Davidson, S., Frew, J.: Why data citation is a computational problem. Commun. ACM **59**(9), 50–57 (2016). https://doi.org/10.1145/2893181
8. Myers, J., et al.: Towards sustainable curation and preservation. In: Proceedings of the IEEE eScience Conference 2015, pp. 526–535 (2016)
9. Belhajjame, K., et al.: A suite of ontologies for preserving workflow-centric research objects. J. Web Semant. **32**, 16–42 (2015)

10. Martin, P., Remy, L., Theodoridou, M., Jeffery, K., 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
11. Bordawekar, R., Blainey, B., Apte, C.: Analysing analytics. SIGMOD Rec. **42**, 4 (2014)
12. Filgueira, R., Krause, A., Atkinson, M., Klampano, I.: dispel4py: a python framework for data-intensive scientific computing. IJHPCA **31**, 316–334 (2016)
13. Ahanach, E., Koulouzis, S., Zhao, Z.: Contextual linking between workflow provenance and system performance logs. In: 15th IEEE International Conference on e-Science, San Diego, US (2019). http://doi.org/10.1109/eScience.2019.00093
14. 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
15. Koulouzis, S., et al.: Time-critical data management in clouds: challenges and a Dynamic Real-time Infrastructure Planner (DRIP) solution. Concurr. Comput. Pract. Exp. (2019). https://doi.org/10.1002/cpe.5269
16. de la Hidalga, A.N., et al.: The ENVRI Reference Model (ENVRI RM) version 2.2 (2017). http://doi.org/10.5281/zenodo.1050349