ICT Infrastructures for Environmental and Earth Sciences

Jeffery, K.; Pursula, A.; Zhao, Z.

DOI
10.1007/978-3-030-52829-4_2

Publication date
2020

Document Version
Final published version

Published in
Towards Interoperable Research Infrastructures for Environmental and Earth Sciences

ICT Infrastructures for Environmental and Earth Sciences

Keith Jeffery[1](✉), Antti Pursula[2], and Zhiming Zhao[3]

[1] Keith G Jeffery Consultants, Faringdon, UK
keith.jeffery@keithgjefferyconsultants.co.uk
[2] CSC - IT Center for Science, Espoo, Finland
antti.pursula@csc.fi
[3] Multiscale Networked Systems, University of Amsterdam, 1098XH Amsterdam, The Netherlands
z.zhao@uva.nl

Abstract. E-Infrastructures play an increasingly important part in the provision of digital services to environmental researchers and other users. The availability of reliable networks, storage facilities, high performance and high throughput computers and associated middleware and services to ease their utilisation all contribute to enabling research and its exploitation. Their relevance, possible use and utilisation to date are described.

Keywords: Infrastructure · Open Science · Networking · Computers · Cloud computing

1 Introduction

To tackle the scientific challenges discussed in the previous chapter, researchers need access to sophisticated *research support environments* that enable efficient discovery, access, interoperation and re-use of the data, tools, etc. available for advanced data science and provide a platform for the integration of all resources into cohesive observational, experimental and simulation investigations with replicable workflows. Examining current initiatives in Europe and beyond, we have identified three main types of research support environment [1]:

e-Infrastructures. Unified computing, storage and network infrastructures provided via initiatives such as EGI[1], GEANT[2], and EUDAT[3]. The e-Infrastructure providers manage the *service lifecycle* of computing, storage and network resources, and enable research communities to provision dedicated infrastructure and to manage persistent services and their underlying storage, data processing and networking requirements.

[1] http://www.egi.eu/.
[2] http://www.geant.org/.
[3] http://www.eudat.eu/.

Z. Zhao and M. Hellström (Eds.): Towards Interoperable Research Infrastructures for Environmental and Earth Sciences, LNCS 12003, pp. 17–29, 2020.
https://doi.org/10.1007/978-3-030-52829-4_2

Public e-Infrastructures typically offer their services based on service-level agreements (SLAs) established at the institutional level or negotiated with specific groups [5]. Such services are now predominantly Cloud-based, using virtual machines or containers that can be easily migrated and scaled across clusters of generic hardware.

Research Infrastructures (RIs). Dedicated data infrastructures constructed by specific scientific communities for combining scientific data collections with integrated services for accessing, searching and processing research data within specific scientific domains; examples include the Integrated Carbon Observation System (ICOS)[4] for carbon monitoring in atmosphere, ecosystems and marine environments, the European Plate Observing System (EPOS)[5] for solid Earth science and Euro-Argo[6] for collecting environmental observations from large-scale deployments of robotic floats in the world's oceans. RIs play a key role in the *research data lifecycle*, providing standard policies, protocols and best practices for the acquisition, curation, publication, processing and further usage of research data and other assets such as tools and simulation/modelling platforms. They typically work closely with (or effectively subsume) individual data centres dedicated to research data, sensor networks, laboratories and experimental sites.

Virtual Research Environments (VREs). Platforms providing user-centric support for discovering and selecting data and software services from different sources, and composing and executing application workflows [3], also referred to as Virtual Laboratories [2] or Science Gateways [3]. Examples include VRE4EIC[7], D4Science[8] and EVER-EST[9]. VREs play a direct role in the *activity lifecycle* of research activities performed by scientists, for example, the planning of experiments, search and discovery of resources from different sources (notably including RIs), integration of services into cohesive workflows and collaboration with other scientists [4]. Graphical environments, workflow management systems, and data analytics tools are typical components of such environments.

While the roles and functions of these different kinds of environment may substantially overlap, none individually fulfil all the requirements of data-centric research; in practice, all these types of research support environment must be tightly integrated (and their overlapping functions reconciled and duly delegated). In particular, e-infrastructures focus on generic ICT (Information and Communication Technologies) resources (e.g. computing or networking), RIs manage data and services focused on specific scientific domains, and VREs support the lifecycle of specific research activities. Although, as already noted, the boundaries between these environments are not always entirely clear (often sharing services for infrastructure and data management), collectively they represent an important trend in many international research and development projects. Figure 1 shows the abstract logical relationship between e-infrastructures, RIs and VRE.

[4] https://www.icos-ri.eu/.
[5] https://www.epos-ip.org/.
[6] http://www.euro-argo.eu/.
[7] http://www.vre4eic.eu/.
[8] https://www.d4science.org/.
[9] https://ever-est.eu/.

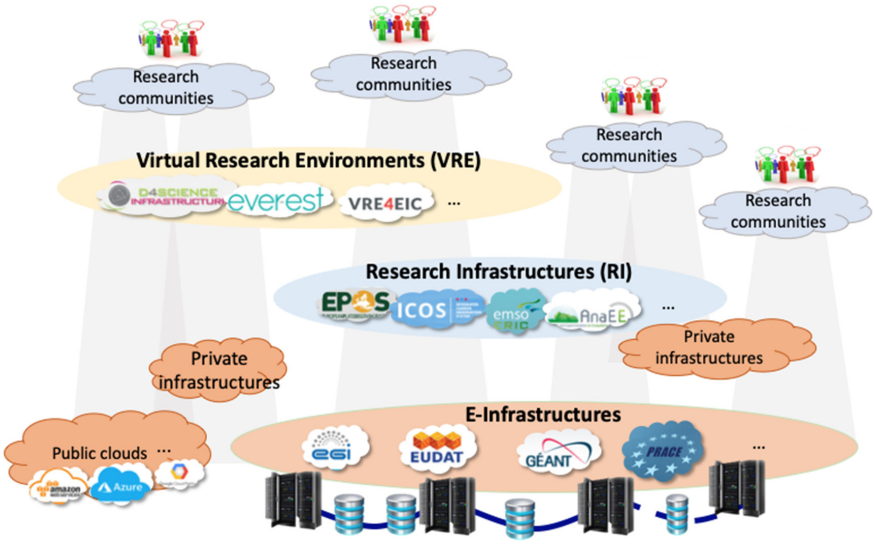

Fig. 1. A layered view of the different kinds of research support environment used by research communities.

Like other domains of research, environmental science has progressively adopted ICT. Perhaps more than other domains, environmental science has complexity because it encompasses observational, experimental and modelling/simulation methods across complex natural systems which have a past, a present and a predicted future. The RIs in environmental and Earth sciences commonly have their own ICT infrastructures but increasingly utilise e-Infrastructures external to the RI and shared commonly among multiple domains of research. This chapter characterises those e-Infrastructures and places them within the ENVRI framework.

In this chapter, we will introduce some typical examples of e-infrastructures. Based on those low-level ICT technologies and infrastructures, we will discuss the research infrastructures and Virtual Research Environments in the later chapters.

2 The e-Infrastructures

This section outlines the e-Infrastructures of relevance to ENVRI, their characteristics and offerings and how they have been used by RIs in ENVRI.

2.1 GEANT

GEANT[10] is the pan-European network for research and education and links seamlessly with other continental networks to form an international communications infrastructure. GEANT was formed by connecting the NRENs (National Research and Education Networks) and has since provided a high speed (100 Gb/s), reliable (100%) network beyond the capabilities of commercial suppliers in order to support leading-edge academic activity.

The RIs of ENVRI depend totally on GEANT for connectivity to the world outside of the RI. In some cases, where RIs have multiple institutions or facilities within them dispersed geographically, they depend on GEANT for communications within the RI. The RIs in ENVRI use services over GEANT for accessing computer systems, using WWW (World Wide Web) facilities, for email and teleconferencing, for file transfer, for control of instruments for observation and experiments and more.

2.2 EGI

Arising from a European Grid Initiative (sharing resources across Europe and beyond) EGI[11] is a federation and not-for-profit organisation providing virtualised access to multiple e-Infrastructures providing computing resources (through HTC and Cloud computing) and storage (online and archival), and services for data processing (i.e., Jupyter Notebook), data management (i.e., Datahub), and AAI (i.e., Check-in).

Various RIs in ENVRI have used EGI facilities to provide computing and storage resources beyond the capability of the RI itself. EGI staff involved in ENVRIplus have supported joint pilot projects with RIs to demonstrate the capabilities of the EGI facilities.

2.3 EUDAT

EUDAT[12] offers an e-Infrastructure for storage and associated services. The EUDAT CDI (Collaborative Data Infrastructure) is essentially a European e-infrastructure of integrated data services and resources to support research. This infrastructure and its services have been developed in close collaboration with over 50 research communities spanning across many different scientific disciplines and involved at all stages of the design process. The establishment of the EUDAT CDI is timely with the imminent realisation of the European Open Science Cloud (EOSC)[13], which aims to offer open and seamless services for storage, management, analysis and re-use of research data, across borders and scientific disciplines.

EUDAT services include B2FIND for searching a catalogue of available datasets described by CKAN[14] with its metadata schema[15] (although commonly enlarged);

[10] https://www.geant.org/Networks.
[11] http://www.egi.eu/.
[12] https://eudat.eu/.
[13] https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud.
[14] https://ckan.org/.
[15] https://ckan.org/portfolio/metadata/.

B2SHARE and B2DROP for data deposit and B2ACCESS for access control. B2STAGE transfers a dataset to local storage for processing while B2SAFE provides storage and curation facilities.

The EUDAT services for data management are utilised to a various extent by a number of ENV RIs, including eLTER, ICOS and Euro-Argo. The capabilities of B2FIND were demonstrated in ENVRI with a catalogue utilising the CKAN metadata schema (extended) providing access to datasets. Some of the pilot projects performed jointly with EGI staff within ENVRIplus utilised EUDAT, for instance the B2SAFE data storage used by Euro-Argo was extended with a EUDAT Data subscription functionality in an ENVRIPlus use case (ref. Chapter 16).

2.4 PRACE

PRACE[16] (Partnership for Advanced Computing in Europe) is an e-Infrastructure consisting of supercomputer facilities in Europe. The computer systems and their operations accessible through PRACE are provided by 5 PRACE members (BSC representing Spain, CINECA representing Italy, ETH Zurich/CSCS representing Switzerland, GCS representing Germany and GENCI representing France). Four hosting members (France, Germany, Italy, and Spain) secured funding for the initial period from 2010 to 2015. In 2016 a fifth Hosting Member, ETH Zurich/CSCS (Switzerland) opened its system via the PRACE Peer Review Process to researchers from academia and industry. In pace with the needs of the scientific communities and technical developments, systems deployed by PRACE are continuously updated and upgraded to be at the apex of HPC technology. Applications to use PRACE are peer-reviewed to provide project access for typically 3 years. Preparatory projects (to prepare for project access) are supported.

Individual researchers from various RIs in ENVRI have used PRACE facilities for particular research activities but there is no wholesale use of PRACE by ENVRI RIs at present.

2.5 OpenAIRE

OpenAIRE[17] has grown through a series of project phases funded by the European Commission: from the DRIVER projects to link Europe's scholarly publication repository infrastructure, to the first OpenAIRE project aimed to assist the EC in implementing its initial pilot for Open Access (OA) to publications, and, through several further phases which have extended and consolidated the OpenAIRE mission to implement Open Science policies. OpenAIRE has been providing the standards and services (e.g. harvesting, retrieval) to allow a catalogue of research assets to be built and used based on CERIF[18] under an agreement with euroCRIS[19]. CERIF provides the fully connected graph model with base entities and linking (relationship) entities with the role and temporal duration required for describing accurately the word of research.

[16] http://www.prace-ri.eu/.
[17] https://www.openaire.eu/.
[18] https://www.eurocris.org/cerif/main-features-cerif.
[19] https://www.eurocris.org/.

Many researchers in RIs within ENVRI use OpenAIRE directly for searching for relevant publications or other research assets (e.g. datasets) or - indirectly via their institutional repository - through harvesting of metadata on scholarly publications or other research assets to the catalogue. OpenAIRE has another lesson for ENVRI: because of the heterogeneity of metadata formats in the various repositories of research assets, the project discovered that simple metadata schemes were inadequate and chose to use the rich metadata model of CERIF to allow ingestion of the various heterogeneous metadata models describing the distributed institutional assets.

2.6 EOSC

EOSC[20] (European Open Science Cloud) is an initiative funded by the EC to provide a 'commons' for networking, computing resources, storage, services and assets useful to research, industry and citizens. Feasibility has been demonstrated through the EOSC Pilot[21]. EOSC is still under construction and is centred around the EOSC Hub[22] but there are also other more recent projects for constructing the EOSC such as EOSC Secretariat supporting the EOSC governance as well as facilitating a number of European working groups. The facilities are provided by EGI, EUDAT, Indigo Data Cloud[23] and OpenAIRE utilising GEANT.

RIs in ENVRI have participated, first in some joint work with EGI and then in the EOSC Pilot where work was concentrated on metadata and interoperability of data and services. Currently, ENVRI RIs interact with building the EOSC through the ENVRI-FAIR project [8]. A key point about EOSC is that it is built around the concept of services and provides a catalogue of services. Most ENVRI RIs provide catalogues of datasets and so there is a mismatch. Uniquely, EPOS within ENVRI designed and built its catalogue of assets to encompass services, datasets, data products, workflows, software modules, equipment and other research assets, concentrating first on services to align with the evolving EOSC. Furthermore, EPOS uses CERIF and so has a rich metadata format allowing interconversion with less rich metadata formats and also ensuring compatibility with OpenAIRE.

2.7 Sensor Networks

Sensor networks are essential for observation in environmental science. Modern networks are digital with local processing power - sometimes referred to as Fog or Edge Cloud Computing. Many modern sensors can be configured remotely to detect one or more physical attributes (e.g. temperature, pressure, salinity and pH) and to adjust precision and accuracy. By their nature, many sensor networks are specific to a particular RI within ENVRI but some sensor networks are shared among several RIs.

A specialised kind of sensor is earth observation satellites. In this case, the RIs in ENVRI receive data products particularly images in various wavebands (after sensing,

[20] https://www.eosc-portal.eu/.
[21] https://eoscpilot.eu/.
[22] https://www.eosc-hub.eu/.
[23] https://www.indigo-datacloud.eu/.

calibration and any necessary corrections and further processing) from agencies such as ESA (European Space Agency). Many RIs in ENVRI use such services. Similarly, geodesy services utilising satellites including GPS (Global Positioning System) provide information on surface elevation changes. This is used by several RIs in ENVRI from generating 3-D topographic models to detecting earth movements e.g. earthquakes.

2.8 Laboratory Equipment

RIs in ENVRI use laboratory equipment for a variety of purposes from chemical analysis and work on DNA to flumes for hydrological studies and pressure cells for rock mechanics. By their nature, they tend to be specialised to a particular RI although it is possible to utilise external commercial services for some equipment use where the equipment cost is not justified by the amount of likely use. The equipment is usually commercially produced with proprietary formats for data and metadata recording the experiment. Increasingly the equipment has digital capabilities for output and also increasingly for input to control the equipment during the experiment. This opens the possibility of a researcher sending a sample to a particular laboratory and both monitoring/controlling the experiment and collecting the experimental data remotely.

RIs in ENVRI have a large variety of equipment utilised within each RI.

2.9 Computing

RIs in ENVRI have computing equipment within their institutions, and in addition, they may be utilising local or national computing centres for this. These are used for data collection and processing. There is little sharing of such facilities among RIs, nor much sharing of software or even best practice in the use of such equipment across RIs. It is to be hoped that progressively the RIs in ENVRI will appreciate the benefits of shared best practice and software (decreasing costs, increasing professionalism, permitting interoperability) and even sharing of computing resources so that idle computing capacity may be utilised. However, it may be that the cost of data transfer and potential security/privacy risks outweigh the cost savings.

3 Access to the e-Infrastructures

The e-Infrastructures are to be used for research, education and wealth creation and - in the case of ENVRI - there is an opportunity to take advantage of the facilities. However, access to e-Infrastructures requires passing some controls.

3.1 AAAI

AAAI (Authentication, Authorisation, Accounting Infrastructure) refers to the process whereby an end-user gains access to computing and other digital facilities. Typically, from a non-commercial background, a researcher applies to the local institution which authenticates her manually (usually with an email address and password) which in turn

provides access with online authentication via EduGAIN[24] to GEANT and thence - subject to authorisations - to other e-Infrastructures (federated identity management). The authorisation is more complex and is e-Infrastructure-specific (or, for that matter, RI-specific). The RI defines policy and this is then enacted. If the policy is for total open access no authorisation is required although accounting will be required to record accesses as needed by GDPR (General Data Protection Regulation) [6]. Usually, the RI catalogue provides the relationship (authorisation) between an authenticated user and research assets; the relationship being the actions authorised within a role (e.g. execute, read, update, write, and delete) and referred to as RBAC (Role-based access control) [7]. The access may be temporally limited e.g. to ensure no overuse of computing resource or to embargo access to a research asset while the lead researcher(s) publish based on that asset. This is temporally bound RBAC.

3.2 TNA

TNA (Trans-National Access) is a scheme designed to allow researchers from one RI or community to utilise equipment at another. The TNA process is essentially matching a researcher requirement to perform an observation or experiment with a RI that has the appropriate equipment available. It may be compared to hotel reservation systems, although the specifications tend to be more complex and the governance and funding arrangements need to be agreed - ideally generally and in advance. It is expected that the use of the equipment is acknowledged and - in some cases - that publications based on the results are joint between the researcher and staff at the RI owning the equipment, especially if the equipment requires complex and expensive set-up.

Within ENVRI there appears to be little use of TNA. In EPOS a TNA system - accessed from the EPOS portal - is being implemented (currently being tested) to try to optimise the use of expensive laboratory equipment.

4 Aspects of Future Infrastructure

The technologies are evolving constantly. Here some significant developments are outlined and their importance to ENVRI estimated.

4.1 Smart Networks

Smart Networks, commonly known as SCN (Software Controlled Networks) are becoming a reality increasingly. They have the ability to manage the available bandwidth on a network segment to obtain maximum throughput together with recording monitoring information to enable dynamic improvements. This is important for RIs in ENVRI, especially for data collection from observations (sensor networks) or experiments (equipment) where there may be very high data rates.

[24] https://edugain.org/.

4.2 Cloud Dynamic Resource Allocation

Cloud computing virtualises computing resources so that the end-user neither knows nor cares where their computing is being done. Building upon the concept of GRIDs developed through an EC Expert Group 2000–2006[25] Cloud Computing was considered by another EC Expert Group[26]. The major obstacles to Cloud Computing were identified as (a) security, privacy and trust; (b) availability; (c) lock-in to one supplier. Despite some sensational difficulties over availability (when a large computer centre was out of action due to a security attack or power outage) in general (a) and (b) have been overcome. (c) was overcome by techniques to describe an application workflow such that it (or semi-independent components of it) could be deployed by a controlling middleware across one or more Cloud suppliers using VMs (Virtual Machines). Further work led to the optimisation of deployments depending on cost, elapsed time, Cloud supplier computer characteristics (e.g. kind of processor). All of this depended on containerisation – using containers (typically Docker[27]) and a container management environment (with scaling and deployment) such as Kubernetes[28]. Various ENVRI RIs have been experimenting with using such computing environments and it is expected that such architectures will become prevalent in the future.

5 Looking Backward and Forward

The RIs within ENVRI - like all RIs within the ESFRI family - have been on a journey over the last few years, increasing their capabilities and knowledge and adapting to the opportunities provided by the new, emerging technologies and the ever-increasingly ambitious requirements of the researchers and other users. Here we assess the journey during the ENVRIplus project and suggest some future projections.

5.1 Shared Experience

There has been much sharing of experience during ENVRIplus. Now each RI in ENVRI has an appreciation of the way each other RI has developed its ICT. There has been an increasing realisation that there are opportunities for sharing of more tangible assets such as software and leading to the end-goal of interoperability so that a researcher in one domain can utilise the assets of other domains to form a more comprehensive understanding of the environment.

5.2 Shared Best Practice

Different ENVRI RIs started at different stages of development of their ICT. There was an expectation that associated best practice could be shared to improve the offerings and utilisation of each RI by cross-adoption among them of appropriate better techniques.

[25] https://www.ercim.eu/publication/Ercim_News/enw66/jeffery.html.
[26] https://ercim-news.ercim.eu/en80/es/the-future-of-cloud-computing.
[27] https://www.docker.com/.
[28] https://kubernetes.io/.

In some areas this has been demonstrated: there is much more awareness of the need for curation of assets, for example, using the DCC (Digital Curation Centre) model and DMP (Data Management Plan) template. Similarly, awareness has been raised in the areas of use of PIDs (Permanent Identifiers), Citation and rich metadata including for provenance [9]. It is to be expected that in future further convergence of best practice – but specialised for each RI domain – will occur leading to greater opportunities for sharing and an overall raising of standards of research support in all RIs.

5.3 Shared Sensor Networks

Some ENVRI RIs share already sensor networks especially when the equipment is expensive or located remotely – examples are some oceanic instruments whether associated with a particular research vessel cruise or (semi-)permanently positioned. Several RIs use data products derived from satellite sensors. It is to be anticipated that such shared use will increase in the future as the costs of deploying the sensors increases, the sophistication of the network control (through Fog/Edge Cloud computing) - allowing autonomic operation - increases and the research requirements demand more shared use of multiple sensor networks to produce a multidomain environmental analysis.

5.4 Shared Equipment

Some RIs are or have institutions which own and use particular experimental equipment. While some equipment is inexpensive and there is no real advantage in sharing, in other cases not only is the equipment expensive but the experienced technicians and researchers needed to operate the equipment are expensive. Therefore, there is merit in sharing. While in some cases a service may be offered such that a sample may be sent, analysed at the equipment and the results returned digitally in other cases it is recommended that the researcher attend the equipment themselves to fully understand the capabilities and limitations (including accuracy, precision and calibration) of the equipment. It is to be expected that there is more equipment sharing in the future; the 'remote service' kind being more common than the 'attend and operate the equipment' kind. In EPOS, for example, a prototype system for TNA (Trans-National Access) is being tested where the researcher request for access to equipment is matched with suitable equipment availability and the agreement facilitated.

5.5 Shared RI Computing

The ENVRI RIs have their own computing equipment, usually used as servers to perform computing tasks and to provide data storage. Some have their computing resources operating as a cluster and a few have utilised Cloud middleware to provide an in-house private Cloud service. It is true generally that these separate, distributed and distinct computing resources are not utilised fully. If they could be coupled together, appropriate middleware installed and canonical systems for security, privacy, trust and resource management introduced then there would be two benefits: (1) each RI would have available more compute and storage capacity; (2) the assets of any RI would become more easily

interoperable since the asses would be virtualised in the Cloud environment. However, RIs would also perceive disbenefits: (1) there is clearly a security risk in opening up computing resources previously private to a wider networking environment; (2) local management of the resources of a single RI would no longer necessarily have precedence so an urgent task may not be allowed to run immediately. This could be particularly important if real-time data is being streamed to the RI computing centre. (1) could be partially overcome by 'sandboxing' any executable software deployed to computing resources other than the RI of origin. (2) could be overcome by system management overrides in the resource allocation system. This may be a possible route forward for the future but would require a degree of cooperation in governance as yet not foreseen.

5.6 Shared External Computing

As indicated earlier, several ENVRI RIs have experimented with using external services such as those provided by EGI and PRACE in order to gain more computing power than available at the local RI computer centre. Similarly, some have experimented with using EUDAT for data storage. Many have used OpenAIRE for its curation and provenance capabilities (a central canonical catalogue pointing to open repositories) for documents and increasingly for datasets. During ENVRIplus no consistent policy shared across the ENVRI RIs emerged. However, the emergence of the EOSC concept provides an architectural basis for the utilisation of external resources since the above external e-Is are collected under that umbrella. Various ENVRI RIs were involved in the EOSC Pilot project particularly considering interoperation sanctioned by conversion of heterogeneous metadata schemas to a canonical rich metadata format. It is expected that during the ENVRI-FAIR project that a coherent policy for external access covering governance, sustainability and FAIR principles[29] as well as technical architecture based on a common, logically-centralised rich metadata catalogue will be adopted by ENVRI RIs.

5.7 Shared Datasets

ENVRIplus concentrated on datasets as primary assets of the RIs. Some scientific use cases had requirements for datasets from several RIs and thus some datasets were shared. However, the datasets usually required some management and manipulation in order to make them reusable: this usually involved unit conversions or adjustment of spatial coordinates. The overall aim of ENVRI is to make datasets (and other assets) shareable so that more comprehensive environmental analyses may be achieved. In the future, it is to be anticipated that datasets described by rich metadata (as recommended by FAIR and a being parameterised by the FAIR Data Maturity Working Group of RDA[30]) will be shareable.

5.8 Shared Workflows

Many workflows are – by their nature – specific not only to a domain but to a particular part of a subdomain. However, for purposes of reproducibility of research results, it is

[29] https://www.force11.org/group/fairgroup/fairprinciples.

[30] https://www.rd-alliance.org/groups/fair-data-maturity-model-wg.

important that workflows be stored, characterised by rich metadata and made available. Furthermore, taking a pre-existing and available workflow and modifying it for a new purpose may save much research effort. This depends – of course – on the assets utilised by the workflow being FAIR. Once a workflow is shared comes the challenge of workflow deployment. This is where the work on interoperable multi-Clouds described above becomes especially valuable.

5.9 Shared Software

One of the aims of ENVRIplus was to share software – either pre-existing at one or more RIs or developed within the project. The software envisaged was of two types: (1) software that every RI needed to manage its assets including software for curation, cataloguing, asset access, provenance; (2) software required to interoperate data across RIs. During the project both types were specified formally with the RM (Reference Model) but (a) no existing software met exactly the specification; (b) the project did not have the resources to develop the required software.

5.10 Shared Services

EPOS made the decision to first catalogue services as assets rather than datasets (although datasets, equipment and other assets are being catalogued now). This was for several reasons: (1) it was clear that the proposed EOSC was going to be based on services and EPOS wished to have EOSC interoperability; (2) by offering a service, a provider implicitly also offers (a) access to the dataset(s) utilised; (b) to the computing resources required to execute the service; (c) data management services to reduce the dataset(s) only to those records that meet the parameters input by the user. Progressively some ENVRI RIs offer services through their portals as well as access to datasets for download. Longer-term the concept of data download (analogous to using a library catalogue card to find a book then taking it home to read) may become unviable since datasets are growing larger and network speeds are not increasing at the equivalent rate. Hence the concept of user-controlled (or control by software acting on behalf of the user) data management at a remote RI becomes necessary – and this implies access through a service.

5.11 Interoperation - Shared Metadata (FAIR)

The vision of ENVRI is that a researcher, policymaker, commercial user or citizen at any location can ‘see’ through a catalogue a homogeneous view over the heterogeneous assets available at the ENVRI RI sites. The optimal way to achieve this is through homogenised rich metadata (as discussed in Chapter 8) derived from the heterogeneous local metadata standards utilised at each RI. Mechanisms to achieve this matching and mapping of metadata schemas are discussed in Chapter 8, as is the need for and benefits of rich metadata. In order for the assets to be FAIR, they need to meet certain standards or achieve appropriate scores against parameters currently being defined by the RDA FAIR Data Maturity Working Group. The current ENVRI RIs are all - to some extent- FAIR but few reach the more advanced aspects of FAIR, The ENVRI-FAIR project should assist in improving the FAIRness of assets in all ENVRI 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. Koulouzis, S., et al.: Time critical data management in clouds: challenges and a Dynamic Infrastructure Planner (DRIP) solution. Concurr. Comput. Pract. Exp. e5269 (2019). https://doi.org/10.1002/cpe.5269
2. 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. Int. J. Future Gen. Comput. Syst. **101**, 1–13 (2019). https://doi.org/10.1016/j.future.2019.05.076
3. Miller, M.A., Pfeiffer, W., Schwartz, T.: The CIPRES science gateway: enabling high-impact science for phylogenetics researchers with limited resources. In: Proceedings of the 1st Conference of the Extreme Science and Engineering Discovery Environment: Bridging from the Extreme to the Campus and Beyond, Chicago, IL (2012)
4. Remy, L., et al.: Building an integrated enhanced virtual research environment metadata catalogue. J. Electron. Libr. (2019). https://zenodo.org/record/3497056
5. Skene, J., Emmerich, W., Raimondi, F.: Service-level agreements for electronic services. IEEE Trans. Softw. Eng. **36**(02), 288–304 (2010)
6. Tene, O., Evans, K., Gencarelli, B., Maldoff, G., Zanfir-Fortuna, G.: GDPR at year one: enter the designers and engineers. IEEE Secur. Priv. **17**(06), 7–9 (2019)
7. Ghafoor, A., Joshi, J., Latif, U., Bertino, E.: A generalized temporal role-based access control model. IEEE Trans. Knowl. Data Eng. **17**, 4–23 (2005)
8. 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, pp. 277–280. IEEE (2019). https://doi.org/10.1109/escience.2019.00038. https://zenodo.org/record/3462816
9. Wofford, M., Boscoe, B., Borgman, C., Pasquetto, I., Golshan, M.: Jupyter notebooks as discovery mechanisms for open science: citation practices in the astronomy community. Comput. Sci. Eng. **22**(01), 5–15 (2020)