D9.3
Serving key data service stakeholders and policy initiatives

WORK PACKAGE 9 – Service Validation and Deployment

LEADING BENEFICIARY: LUND UNIVERSITY

Author(s):	Beneficiary/Institution
Margareta Hellström (lead) & Alex Vermeulen	LU (ICOS, Lund University)
Yin Chen & Baptiste Grénier	EGI.eu (EGI Foundation)
Markus Stocker & Robert Huber	UniHB (Marum, University of Bremen)
Ingemar Häggström & Carl-Fredrik Enell	EISCAT (EISCAT Scientific Association)
Glenn Judeau & Thierry Carval	IFREMER (EuroArgo)
Leonardo Candela	CNR (ACTRIS)
Jani Heikkinen & Chris Ariyo	CSC (EUDAT)
Domenico Vitale	UNITUS (ICOS, University of Tuscia)

Accepted by: Zhiming Zhao (Theme 2 leader)

Deliverable type: REPORT

Dissemination level: PUBLIC

Deliverable due date: 31.10.2017/M30

Actual Date of Submission: 31.10.2017/M30

A document of ENVRI[plus] project - www.envri.eu/envriplus

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 654182

ABSTRACT

This deliverable reports the group efforts of Working Package 9 Task T9.2 on serving key data service stakeholders and policy initiatives during M13 to M30.

Project internal reviewer(s):

Project internal reviewer(s):	Beneficiary/Institution
Laura Beranzoli	INGV (EMSO)

Document history:

Date	Version
15.7.2017	Outline available for comments
19.9.2017	First draft
6.10.2017	Version 1.0, ready for internal review
26.10.2017	Updated version, incorporating internal reviewer comments, ready for WP and Theme leader review
30.10.2017	Version incorporating Theme leader comments, ready for project office review
31.10.2017	Accepted by project office and submitted

DOCUMENT AMENDMENT PROCEDURE

Amendments, comments and suggestions should be sent to Margareta Hellström (lead author; margareta.hellstrom@nateko.lu.se) and Alex Vermeulen (Task 9.2 leader; alex.vermeulen@icos-ri.eu).

TERMINOLOGY

Acronyms used in this report are briefly explained in **Appendix A**. In addition, the project glossary is found at https://envriplus.manageprojects.com/s/text-documents/LFCMXHHCwS5hh.

ENVRIplus PROJECT SUMMARY

ENVRIplus[1] is a Horizon 2020 project bringing together environmental and Earth system research infrastructures (RIs), projects and networks together with technical specialist partners to create a more coherent, interdisciplinary and interoperable cluster of Environmental Research Infrastructures across Europe. It is driven by three overarching goals: 1) promoting cross-fertilization between infrastructures, 2) implementing innovative concepts and devices across RIs, and 3) facilitating research and innovation in the field of environment for an increasing number of users outside the RIs.

ENVRIplus aligns its activities to a core strategic plan where sharing multi-disciplinary expertise will be most effective. The project aims to improve Earth observation monitoring systems and strategies, including actions to improve harmonization and innovation, and generate common solutions to many shared information technology and data related challenges. It also seeks to harmonise policies for access and provide strategies for knowledge transfer amongst RIs. ENVRIplus develops guidelines to enhance transdisciplinary use of data and data-products supported by applied use-cases involving RIs from different domains. The project coordinates actions to improve communication and cooperation, addressing Environmental RIs at all levels, from management to end-users, implementing RI-staff exchange programs, generating material for RI personnel, and proposing common strategic developments and actions for enhancing services to users and evaluating the socio-economic impacts.

ENVRIplus is expected to facilitate structuration and improve quality of services offered both within single RIs and at the pan-RI level. It promotes efficient and multi-disciplinary research offering new opportunities to users, new tools to RI managers and new communication strategies for environmental RI communities. The resulting solutions, services and other project outcomes are made available to all environmental RI initiatives, thus contributing to the development of a coherent European RI ecosystem.

[1] http://www.envriplus.eu/

EXECUTIVE SUMMARY

Environmental RIs are part of a complex landscape of stakeholders which includes both large actors like regional and global Earth Observation systems, as well as a diverse group of end user communities from a wide range of domains and disciplines. This creates a need to define common strategies and procedures in order to optimise contact interfaces, data and metadata streams, and overall cooperation with these key data users.

To facilitate both their own internal workflows, as well as to assist their end users, ENVRIplus partners are actively developing a wide palette of services and tools for data collection, curation, analysis and dissemination. Work Package 9 (WP9) is concerned with supporting the development, testing and deployment of these services and tools, by 1) coordinating agile development activities in the form of "use cases" (in addition to those driven by WPs 5-8); 2) identifying generic mechanisms for data access, replication and transfer, between end users and environmental RIs, that fulfil needs throughout the research data lifecycle; 3) improving the usability of developed services in their intended environments through e.g. quality and usability validation; and 4) helping to showcase and advertise the services to stakeholders and other end users, also outside of the environmental and climate science domains.

This deliverable, D9.3 "Serving key data service stakeholders and policy initiative", is mainly concerned with points 2 and 3. Four central "mechanisms" are covered, representing key phases of the environmental research data lifecycle:

1. Near-Real Time data quality control of sensor data
2. Subscription services for data products
3. Using VREs to support user-driven data analysis
4. Optimisation of data and metadata interoperability

For each "aspect", we define the overall ENVRIplus context, including a summary of related agile use case activities (as outlined in D9.1 "Service deployment in computing and internal e-Infrastructures Version1" [Chen 2017]), and identify the main stakeholders and end user communities. We then describe the plans for engaging these key data users in a dialogue aimed at including them in the evaluation and deployment of the aspect-specific services. Finally, we discuss the possible "demonstrators" that will be part of the follow-up deliverable D9.4 "Serving key data service stakeholders and policy initiatives Part 2", due at M46.

TABLE OF CONTENTS

This page left intentionally blank.

1 BACKGROUND

1.1 Goals of Theme 2 ("Data for Science")

Environmental Research Infrastructures are expected to become important pillars not only for supporting their own communities, but also (a) for inter-disciplinary research and (b) for large cross-community and cross-disciplinary initiatives, such as the Global Earth Observation System of Systems (GEOSS). To be able to fulfil these important roles, all data-related activities of the ENVRIPLUS partners need to be well integrated and synchronised. In its turn, this requires common policies, common models for data and architectural models, and shared e-infrastructure to optimise technological implementation, define workflows, and ensure coordination, harmonisation, integration and interoperability of data, applications and other services between the research infrastructure initiatives in the environment thematic area [ENVRIplus 2015b].

1.2 Objectives of Work Package 9 ("Service validation and deployment")

The title of ENVRIplus Work Package 9 is "Service validation and deployment". This work package will collect information about the development, and subsequent deployment, of services taking place within the framework of ENVRIplus. In this context, the term "services" may be seen as covering many different types of entities – ranging from algorithms, software and workflows to fully operational computation or management systems – with the common denominator that they can all be applied by RIs and other stakeholders in order to optimise and facilitate their operations. (See **Chapter 2** for an overview of Theme 2-related services.)

The Theme 2 objectives [ENVRIplus 2015a] with special relevance to WP 9 are:

- to facilitate discovery of software services and their composition;
- to characterise ICT resources (including sensors and detectors) to allow virtualisation of the environment (for instance onto Grid- or Cloud-based platforms) such that data and information management and analysis is optimised in use of resources and energy usage;
- to facilitate the connection of users, composed software services, appropriate data and necessary resources in order to meet end-user requirements;
- to facilitate data discovery and use, and to provide integrated end-user information technology to access heterogeneous data sources;

1.3 The missions of Task 9.2 ("From research to operational")

Task 9.1 ("Validation and integration of developed services") is concerned with promoting development of services driven by science, technology and use cases, and analysing the different developed services from the point of view of their integration into the different e-infrastructures required. Complementary to this, the main mission of Task 9.2 is to focus on tracking the actual usability and operationality of the ENVRIplus services, once these are deployed under real world conditions. (The terms "usability" and "operationality" can be interpreted in many different ways. Of special importance here, is the impact a service or tool has on facilitating data delivery to stakeholders and end users.)

Task 9.2 will not only look at the use cases administered by Task 9.1, but will also look at the use cases being developed in WP 6-8. Of special interest are those services that involve large-scale data delivery, e.g. to major stakeholders. Important issues include:

- Deepening and further developing the integration of RIs in various domains to generic initiatives such as Copernicus Atmospheric Monitoring Services (CAMS), the Global Earth Observation System of Systems (GEOSS), the European Environmental Agency (EEA) and others.
- The joint exploration, together with stakeholders, of problems and issues preventing efficient access to and transfer of (large volumes of) data between the identified key stakeholders and environmental infrastructures
- Designing generic "mechanisms", preferably technology and platform independent, that address the identified problems. Here all phases of the research data lifecycle need to be taken into account
- Reviewing, together with the stakeholders, the designed mechanisms. To broaden the impact towards global interoperability, environmental infrastructures from other regions should also be invited at this stage.

1.4 Setting the scene

1.4.1 The Theme 2 Research Data Management components

As described in, e.g., [ENVRIplus 2015b] and the ENVRIplus deliverable D5.1 [Atkinson 2016], Theme 2 ("Data for Science") is concerned with developing services and models and e-infrastructure that support environmental RIs across all aspects of the research data life cycle – from collection of sensor data all the way to finalised data products.

Theme 2 has chosen to organise its Work Packages according to a model of research data management (RDM) that centres around six "pillars" representing distinct components -- cataloguing, curation, identification & citation, optimization, processing and provenance, see **Figure 1** below. The pillars are connected by 3 "cross beams", representing aspects vital to interoperability such as a system architecture based on a common reference model, and shared services for meta-information exchange. The pillars can be thought of as connectors between on one side the underlying base of e-infrastructures, and on the other all the end users of the envisaged services and systems.

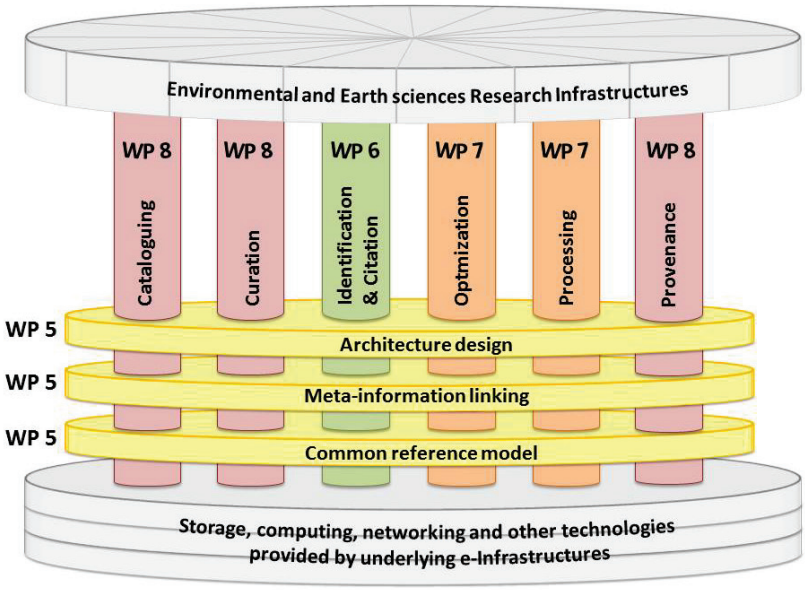

FIGURE 1. THE "PILLARS" AND "CROSS BEAMS" CONNECTING ENVIRONMENTAL AND EARTH SCIENCE RESEARCH INFRASTRUCTURES WITH E-INFRASTRUCTURE PROVIDERS. BASED ON [ZHAO 2015] AND [MARTIN 2017].

1.4.2 The research data lifecycle

Looking closer at the data lifecycle itself (see **Figure 2**), the ENVRI Reference Model [ENVRI RM V2.1 2016] identifies five distinct phases or stages, which can be described as follows:

Acquisition: This is considered to begin upon point of data entry into the RI systems. The acquisition phase as modeled in the ENVRI RM starts from the design of the experiment. Acquisition is typically distributed across networks of observatories and stations. The data acquired is generally assumed to be non-reproducible, being associated with a specific (possibly continuous) event in time and place.

Curation: This phase manages data that is being characterised and supplemented by adding additional information to facilitate identification and retrieval. Supporting activities include quality assessment, annotation, and registering and/or allocating (persistent) identifiers.

Publishing: This step deals with facilitating the access to data that is deemed ready for dissemination, regardless of the amount of analysis or elaboration that may have been undertaken. Supporting activities include providing layers of access control to repositories as well as supporting discovery via catalogues or portals.

Processing: This part covers the derivation of new data products through different types of processing on pre-existing datasets, for example by applying statistical analyses or using data as input for model calculations. Supporting activities include production of syntheses and the derivation of knowledge from information.

Use: this part provides functionalities that manage and track users' activities while supporting the users to conduct their research activities which may result in the creation of new data products. Data 'handled' and produced at this phase are typically user-generated data and communications.

FIGURE 2. THE FIVE STAGES OF THE RESEARCH DATA LIFECYCLE, AS IDENTIFIED IN THE ENVRI REFERENCE MODEL. BASED ON [ENVRI RM V2.1 2016].

2 SERVICE DEPLOYMENT, INTEGRATION AND VALIDATION ACTIVITIES

Through WP9, Theme2 addresses service validation and deployment issues. WP9 is organised into two tasks: T9.1 focuses on service integration and deployment, as well as interfacing e-Infrastructures (e-IRs), while T9.2 works on promoting project products and results, and interfacing RIs. The two tasks are linked and supports to each other. For example, in T9.1, a list of use cases was identified to test Theme2 services in realistic usage scenarios. T9.2 further analyses these use cases from the point of view of RIs, identifying interest areas common across the communities where Theme2-developed services can be part of the solutions. This can make it easy for RIs to understand how to adopt Theme2 services.

However, without validation and evaluation activities, end users will find it difficult to judge the usefulness and applicability of the services. The second activity in T9.1 is therefore to design evaluation metrics to measure the success of use cases implementation. The matrix covers aspects including service usability and impacts. This will provide inputs to T9.2.

T9.1 is also working on setting up the ENVRIplus Service Portfolios, which is a way to manage Theme2 service products. The resulting catalogue[2] (together with evaluation information mentioned above) will be used by T9.2 to communicate and present Theme2 service products to RIs and other end users.

2.1 Service development

In this chapter, we briefly review the on-going service development activities in Theme 2, and how these relate to the research data management models described in **Chapter 1** above. Please refer to Tables 1 and 2 below for more information on agile task force and work package efforts, respectively.

2.1.1 Agile case task forces

Task 9.1 undertakes testing and validation of ENVRIplus service solutions in realistic usage scenarios. As reported in deliverable D9.1 [Chen 2017], the Task has developed a method that jointly collects and identifies appropriate community use cases. Use case proposals were requested from participants, and the received proposals were then reviewed and evaluated to ensure that they reflect strong community interests, have high scientific values, and are well suited to demonstrate the value of ENVRIplus. The process resulted in 13 use cases selected for implementation, as listed in the **Table 1** below (adapted from [Chen 2017]).

2.1.2 Work package-driven efforts

In addition to the agile use cases administered under Task 9.1, there are also a number of "use cases" embedded into the respective Descriptions of Work [ENVRIplus 2015a] of the other ENVRIplus Theme 2 work packages. A majority of these are concerned with the design, development and subsequent deployment of services, and thus should also be covered by Task 9.2 efforts to evaluate, validate and follow up ENVRIplus outcomes. **Table 2** below, ordered according to Work Package, summarises these cases and their related Research Data Management (RDM) components.

[2] https://confluence.egi.eu/display/EC/ENVRIplus+Service+Portfolios

TABLE 1. THE USE CASES IDENTIFIED BY TASK 9.1 AND THEIR RESPECTIVE DATA LIFECYCLE STAGES AND RESEARCH DATA MANAGEMENT (RDM) COMPONENTS. ADAPTED FROM [CHEN 2017].

Code	Case titles	Data lifecycle stage(s)	RDM component(s)
SC_3	How do mosquito-borne diseases emerge and what are the trends?	Use	Processing, Architecture design
TC_2	EuroArgo data subscription service	Curation, Publishing, Processing	Cataloguing, Curation, Identification & citation, Optimisation, Processing, Provenance
TC_4	Sensor registry	Curation	Cataloguing, Curation, Identification & citation, Provenance, Semantic information linking
TC_16	Description of a national marine biodiversity data archive centre	All stages	Semantic information linking, Reference model
IC_1	Dynamic data identification & citation	Curation	Identification & citation
IC_2	Provenance implementation	Curation	Cataloguing, Curation, Identification & citation, Provenance
IC_3	Supporting EISCAT-3D users to reprocess data using users' algorithms	Use, Publishing (accessing the data) and Processing	Optimisation, Processing
IC_9	Quantitative accounting of Open Data use	Publishing	Identification & citation, Provenance
IC_10	Domain extension of existing thesauri	Curation	Semantic information linking
IC_11	Semantic linking framework	Curation	Semantic information linking
IC_12	Implementation of the ENVRI Reference Model for EUFAR and eLTER	All stages	Semantic information linking, Reference model
IC_13	The eddy covariance fluxes of greenhouse gases	Processing	Optimisation, Processing
IC_14	SOS&SSN ontology based data acquisition and NRT technical innovations	Acquisition	Processing

TABLE 2. USE CASES UNDER DEVELOPMENT AS PART OF ENVRIPLUS THEME 2 WORK PACKAGE EFFORTS, AND THEIR RESPECTIVE DATA LIFECYCLE STAGES AND RESEARCH DATA MANAGEMENT (RDM) COMPONENTS. (EXTRACTED FROM THE PROJECT DESCRIPTION-OF-WORK [ENVRIPLUS 2015A].)

WP	Case description	Data lifecycle stage(s)	RDM component(s)
5	Common developments of Euro-Argo and ICOS on CO_2 and pH data	Acquisition, Curation	Meta information linking, Reference model
6	Development of a full data lifecycle model for biological data linked to the standards developed and promoted through GBIF[3]	All stages	Identification & citation, Curation, Provenance
6	Establishing a GBIF Integrated Publishing Toolkit for the publication of data from the Marine Biological Records journal	Curation, Publishing	Curation, Cataloguing, Identification & citation
6	Defining workflows for the marine research community to facilitate the provision of DOIs and the archiving of data in compliance with international standards	Curation, Publishing	Identification & citation, Curation
6	Implementation of a full and integrated DOI and related metadata system for ICOS RI as an example of a distributed RI with very heterogeneous data	All stages	Identification & citation, Cataloguing, Provenance
6	Testing of data citation tracking models, using ICOS RI as a test case.	Curation, Publishing, Production (Use),	Identification & citation, Cataloguing
7	Development of interoperable data processing services	Processing, Use	Processing
7	Development of interoperable process optimisation services	Processing, Use	Optimisation
8	Design and implement prototypes for catalogues of ENVRIplus flagship data products and services	Curation, Publishing	Cataloguing

[3] Global Biodiversity Information Facility, see https://www.gbif.org/.

2.2 Integration

Each use case is defined to combine ENVRIplus Theme2 services with community services to deliver useful functionalities supporting community's need. The implementation of these use cases follows Agile principles[4]. For each use case, an agile team is composed. They respond to test technologies, service integration and deployment. Establishing short-live, agile multi-disciplinary task forces to address specific issues is key to developing the depth of understanding and commitment to make progress.

Most of the agile team members are researchers directly involved in ENVRIplus RIs, with good knowledge of respective community requirements. They bring information and feedback from their RIs. In turn, they can help to promote project results in their communities. T9.2 can make use of these communication channels to reach out ENVRIplus RIs.

2.3 Validation

Although the integration tests are a key ingredient to assure that ENVRIplus service solutions are useful, it is important to measure how use case implementations fulfil end user needs and how they lead to advances over the state-of-the-art. An initial evaluation metric has been developed in T9.1. This T9.1 activity is closely linked to T9.2 that facilitates and encourages the adoption of ENVRIplus Theme2 results. The outputs will assist prospective adopters in their search for appropriate solutions.

Of special importance for Task 9.2 are the validation activities relating to the end user perceptions of service usability and operational status. In order to not only capture initial views and impressions, but also track how well the services are doing over time, it is necessary to set up and maintain long-lasting communications with the various stakeholders and user communities listed below (**Chapter 4**).

In **Chapter 6**, we describe plans to design, implement and operate a web-based tool for collecting evaluation data for the products exposed via the ENVRIplus Service Portfolio[5]. These data will then form the basis for the service validation analysis activities.

3 "MECHANISMS" FOR DATA ACCESS, REPLICATION OR TRANSFER BETWEEN RIs AND USERS

As outlined in its Description of Work [ENVRIplus 2015a], Task 9.2 should "design generic mechanisms to facilitate data access, replication or transfer, in such a way that these are technology and platform independent." With the Reference Model view of the environmental research data lifecycle (see **Chapter 1)** as a starting point, we have identified a set of four central mechanisms, each reflecting a key lifecycle phase:

- Near-Real Time data quality control of sensor data
- Subscription services for data products
- Using VREs to support user-driven data analysis
- Optimisation of data and metadata interoperability

[4] See e.g. https://en.wikipedia.org/wiki/Agile_software_development for an introduction into agile working principles.
[5] See https://confluence.egi.eu/display/EC/ENVRIplus+Service+Portfolios.

For each "aspect", we proceeded to define the overall ENVRIplus context, including a summary of related agile use case activities are connected, and identify the main stakeholders and end user communities. We then describe the plans for engaging these key data users in a dialogue aimed at including them in the evaluation and deployment of the aspect-specific services. The outcomes of this information collection stage, which followed a common template, were the four summary reports included in Appendices B-E.

The following sections summarize the motivation, context and importance of each mechanism.

3.1 Near-Real Time data quality control of sensor data

See **Appendix B** for a comprehensive description of this mechanism.

Motivation:

Clearly, quality control (QC) on observational or other kinds of data is critical in most domains. Checking for quality increases the fitness of data in downstream use, for instance by research communities. Well-designed and reliable automated QC routines are especially important in the Near Real-Time (NRT) context, e.g. for early warning or situation awareness systems.

Context:

In ENVRIplus: This aspect relates to Theme 1, specifically WP 1, WP 2, and WP 3. The aspect is committed to standardizing observational data streams, ideally at the sensor level so that devices directly generate data following one or more suitable standard formats. Since this aspect deals with observational data processing, it furthermore relates to Theme 2, in particular WP 7. Finally, the aspect relates to TC 4 on Sensor Registry. Such registries can provide data about sensors, such as sensor capabilities, which are important to configure QC routines, e.g. for range checks.

Importance:

Outside of the current domain (atmosphere, ecosystems, marine, solid earth)

Clearly, QC - possibly in a NRT context, e.g. for early warning or situation awareness systems - on observational or other kinds of data is critical in most domains. Data quality is important in downstream use, for instance by research communities. However, as the current focus is on environmental RIs, the schema used for standardised data is tailored for observational data, specifically data generated by sensors and, primarily, sensing devices (although easily applicable for human observers). While observations are obviously made in other scientific and research contexts, for instance in the social sciences via surveys and interviews, the current schema would probably have to be adapted to suit the needs of the different context. In contrast to the environmental and earth sciences, other domains may have less of a requirement for NRT processing and rely more on batch processing for QC. The different mode would arguably require a different architecture or at least a different set of technologies. Specifically, in batch processing mode we would probably not include a distributed real-time computation system such as Apache Storm. Rather, the system would surely rely on more classical ETL processing on database systems.

Regionally (Europe) and Globally

It is unclear whether there is a difference between regional and global significance, as the importance of the problem and possible approaches are independent of region. NRT QC is a common theme in RIs, globally. In conjunction with responses to our survey on specific practices among ENVRIplus partners, we have received many critical comments regarding the feasibility of generic NRT QC services. Serious concerns exist among domain specialists. Tacit knowledge is of particular concern as too often information relevant to the parametrization of QC algorithms depends on experience that needs to first be acquired and may not be documented.

3.2 Subscription services for data products

See **Appendix C** for a comprehensive description of this mechanism.

Motivation:

Many research infrastructures aim for providing easy to use and efficient data distribution. Researchers need this feature to collect data for their analyses, models, and studies. Due to accumulating data sources, end-users frequently repeat selection and download of the data they need depending on their criteria to keep their collection up to date. There is a need to avoid this recurring and time consuming task. Moreover, environmental data lifecycle is complex and reflects various stages needed to produce, acquire, cure, publish, and process environmental data. Therefore, the challenging point of this aspect is to enable delivery of frequently updated data to end-users: research infrastructures, scientific communities, and researchers.

Context:

Referring to the Research Data Management model used by ENVRIplus (See **Chapter 1.4**), subscription services for data have strong couplings to a number of components:

- Data and metadata interoperability: the more data and metadata are interoperable, the most data source origins could benefit from this service. By gathered data to common resources we can provide services that could lead to new trans-disciplinary analyses, models, and studies.
- Data identification: the more common data identification methods are used, the more data producers and data consumers could benefit from this service.
- VRE support for user driven analysis: Storing data collections in a convenient space that could be associated with storage and CPU can provide advanced services.

This mechanism could contribute to the development of both a common standard for input streams as well as, in connection to the following mechanism, to enhancements of digital collaborative spaces for researchers and data providers. *In ENVRIplus:* The Euro-Argo data subscription service is covered by TC_2.

Importance

Data subscription services are expected to play an increased role in the future, as the number of data producers and their respective output starts to increase rapidly.

3.3 Using VREs to support user-driven data analysis

See **Appendix D** for a comprehensive description of this mechanism.

Motivation:

For the ENVRIplus RIs, the ultimate goal is to provide quality-checked and calibrated observational data to their user communities. The extent to which an RI supports its users to access data easily and to deliver good and citable research is directly linked with the impact and value of the RI, and thus this issue is given much emphasis when RIs design their ICT systems. In recent years, Virtual Research Environments (VREs) have emerged as an important approach to providing web-based systems to help researchers collaborate.

Context:

Although the implementations of the three use cases are under WP9 efforts, cross-collaboration has been established with other WPs including, WP5, WP7, and WP8. The key service they test is D4Science which is developed by T7.1. Catalogues services/principles developed by WP8 are also integrated, e.g., by use case IC_3. IC_3 also evaluates Reference Model concepts on metadata aspects.

WP7 (T7.1) has set up a VRE for the ENVRIplus community (link to ENVRIplus VRE) using the D4Science platform[6]. Three WP9 use cases (IC_3, IC_13 and SC_3) are evaluating this service. Detailed descriptions of these use cases have been provided in D9.1. In summary, IC_3 aims to support individual scientists from the EISCAT 3D community to process radar data using their own algorithms. In a similar way, SC_3 aims to enable LifeWatch researchers to develop and share data analysis workflows. Likewise, the objective of IC_13 is to optimise the processing of the eddy covariance (EC) data in order to establish a service that can be used by different RIs that use this micrometeorological technique to measure exchanges of greenhouse gases and energy between terrestrial ecosystems and atmosphere.

Importance:

Outside of the current domain (atmosphere, ecosystems, marine, solid earth)

VRE for user-driven data analysis is an commonly requested area not only to environmental scientific community. The implementations from the three use cases can be easily extended to support other domain applications' needs. For example, feedback from agile team IC_3, their implementation would benefit space and solar-terrestrial physics, solar system physics (meteors and asteroids), and astronomy.

In the case of IC_13, the availability of standardised and quality checked EC data product is essential for many purposes: (i) for accurately quantifying the carbon budget, (ii) for better understanding the complex biogeochemical and ecological processes, and (iii) for the verification and tuning of global climate models, mesoscale and weather models, and remote sensing estimates from satellites and aircraft.

[6] See https://www.d4science.org/

Regionally (Europe)

All three use cases come from ESFRI RIs, ICOS, EISCAT_3D and LifeWatch, that have collaborations across Europe. The implementation will benefit their users in the member countries. For example, the investigation in the use case, IC_3, will mainly benefit the European EISCAT member countries, Finland, France, Norway, Sweden, USA and UK, but also data discovery from other countries.

Globally

The collaborations of the involved RIs mentioned in this aspect, indeed beyond Europe level. For example, in the case of IC_3, the global wise collaborations reach China, Japan, Korea, etc. In the case of IC_13, the availability of a computationally efficient VRE designed to the processing of EC data is of interest for ICOS and any other users or RIs using EC data (e.g. AnaEE, eLTER). A standardised processing guarantees homogeneous data more suitable in comparison studies involving different geographical areas.

3.4 Ensuring interoperability of data and metadata

See **Appendix E** for a comprehensive description of this mechanism.

Motivation:

Improving data access is not only about setting up efficient channels for exchanging digital data objects, for example by implementing user-friendly search interfaces and download services. Indeed, without ensuring that all relevant metadata - administrative, provenance-related, descriptive etc. - is just as easily accessible and interpretable by the end user, the data can be considered more or less useless.

The FAIR guiding principles for data, developed by FORCE11, describe how to make published data Findable, Accessible, Interoperable, and Reusable for potential users [FORCE11 2014, Wilkinson 2016]. Findability means making the data possible to find by potential users, e.g. by describing the data with rich metadata. Accessibility means that the data and metadata should be usable in formats that are understandable by humans and machines, e.g. by adding machine-actionable PIDs. Interoperability pertains to using a metadata scheme that is open and well-defined. Reusability means that the metadata are verifiable, machine-readable and can be used to make proper citations [FORCE11 2014].

By following the FAIR principles, the “seven Rs” of data can be fulfilled, i.e. that data are reusable, repurposable, repeatable, reproducible, replayable, referenceable, and respectful [Bechhofer 2013].

Context:

As stated explicitly in the description of Theme 2 [ENVRIplus 2015b], environmental research infrastructures are expected to become important pillars not only for supporting their own communities, but also (a) for inter-disciplinary research and (b) for the European Earth Observation Program Copernicus, which in its turn is a main contributor to the Global Earth Observation System of Systems (GEOSS). To ensure that the ENVRIs can fulfil these roles, it is very important that all data-related activities within ENVRIPLUS are well integrated. This requires not only common policies, models and e-infrastructure, but very importantly also concerted

efforts to 1) re-harmonise and integrate services and 2) enforce interoperability of data, applications and other services.

ENVRIplus: The interoperability of research data is an integral part of the ENVRIplus Theme 2 activities: semantic information linking is covered by WP 5, data identification and persistent identifiers by WP 6, processing and workflows by WP 7 and cataloguing by WP 8. Because of the pervasive nature of interoperability, practically all WP 9 use cases (Table 2) will contribute to the development of FAIRness throughout all data activities of ENVRIplus.

Importance:

For ENVRIplus and its partner RIs

As stated explicitly in the description of Theme 2 [ENVRIplus 2014], environmental research infrastructures are expected to become important pillars not only for supporting their own communities, but also (a) for inter-disciplinary research and (b) for the European Earth Observation Program Copernicus, which in its turn is a main contributor to the Global Earth Observation System of Systems (GEOSS). To ensure that the ENVRIs can fulfil these roles, it is very important that all data-related activities within ENVRIPLUS are well integrated. This requires not only common policies, models and e-infrastructure, but very importantly also concerted efforts to 1) re-harmonise and integrate services and 2) enforce interoperability of data, applications and other services.

Outside of the current domain (atmosphere, ecosystems, marine, solid earth)

Interoperability is important for all scientific domains! In fact, it may be argued that ensuring that data sets are properly described, following standardised vocabularies and information schemata, becomes even more important when data are being (re-)used across domain boundaries, e.g. when a health researcher wants to incorporate data on air pollution as background material for her own clinical studies. If the air quality data isn't well documented, serious mistakes and errors could ensue if the collected information is used in the wrong way by the medical professional. In addition, if the air data is provided only in some proprietary format, the dataset may be inaccessible unless the same software package is made available to the health researcher.

Regionally (Europe) and globally

Funders and regulators in different regions may choose to impose specific rules or recommendations concerning e.g. specific metadata standards that should be adhered to. One example is the INSPIRE Directive from the European Union, which "aims to create a European Union spatial data infrastructure for the purposes of EU environmental policies and policies or activities which may have an impact on the environment" [INSPIRE 2017].

4 DEFINING THE STAKEHOLDER & USER COMMUNITY LANDSCAPE

One of the six objectives of ENVRIplus is to generate common solutions for the challenges faced by environmental and Earth science research initiatives in their use of data and services, including discovery and use, workflow support, data management and user interaction [ENVRIplus 2015b]. All solutions, services, systems and other project outcomes developed by

ENVRIplus should be made available across the entire spectrum of potential end users, ranging from large generic initiatives down to individual research groups, as well as SMEs and other commercial actors. (It should be noted that while this Deliverable is concentrated on non-commercial user communities, this should in no way be interpreted to mean that the commercial sector's needs for data access are in any way less important.)

In the following, we summarize the outcome of the target group analyses performed for each of the four mechanisms outlined in Chapter 3.

4.1 The landscape of generic initiatives

These include COPERNICUS, GEOSS, EEA, IPBES and similar large-scale organisations and projects, with access to high-performance and high-throughput computing resources and substantial storage. They are often engaged in "operational forecasting" or "near real-time" data analysis, which means that they rely on timely access to homogenous data streams following standardised formats and transfer protocols.

Near-Real Time data quality control of sensor data

Large-scale generic initiatives are not the primary users of automated quality control (QC) services; instead they are typically interested only in the outputs from such services. However, the processing systems applied by a large-scale project like Copernicus may well impose specific demands on which QC algorithms and tests are applied to the data that they are able to ingest.

Subscription services for data products

The possibility to receive regular transmissions of data, especially in near-real time, directly from the organisation responsible for the data collection and (pre-)processing, is very important to many large initiatives. The example described in **Appendix C** illustrates how EuroArgo supplies data to the Copernicus Marine Service via a prototype subscription service. Generic initiatives will themselves be interested to operate subscription services for their outputs, based around a trusted repository hosting synchronised versions of their data collections. Such a service may also allow the provision of new features to end users, generating more visibility.

Using VREs to support user-driven data analysis

Large generic initiatives can afford to set up complex and functionality-rich VREs for their own end users, but they can also benefit from accessing and using processing platforms set up by their own input data providers.

Optimisation of data and metadata interoperability

Many of the large "generic" initiatives like Copernicus and GEOSS are operating and/or planning for monitoring and prediction services on a regional or global scale. These services require access to data on e.g. concentrations and fluxes of greenhouse gases, pollutant and aerosol loadings, or other environmental variables from many different sources – some of which are ENVRIplus partners. Successful operation obviously demands that data are easily accessible in a timely manner, that the transfer is expedient and that all data both are traceable, persistently identified, and follow strictly regulated standards with respect to content type.

4.2 Scientific communities

These include RIs and larger research collaborations – both in the ENVRIplus domains themselves and outside. They are mostly interested in finalized data products, and data subscription services can be very useful. Data interoperability is an important issue, as they often want to combine large and complex datasets from different sources in order to feed their own scientific analyses.

Near-Real Time data quality control of sensor data

Currently, the primary stakeholders are surely environmental RIs, specifically those that collect observational data from sensors, sensor networks, monitoring stations. Secondary stakeholders are environmental RIs, possibly virtual infrastructures, which acquire data from other environmental RIs, especially those that acquire data in NRT. RIs in other domains, e.g. life sciences or high-energy physics, that acquire data in real-time are possible stakeholders. In addition to infrastructures that acquire observational data through in-situ monitoring, stakeholders relevant to this aspect also include infrastructures that acquire observational data through remote sensing, specifically satellite based systems for Earth Observation. Such observational data also undergo QC before downstream use.

Subscription services for data products

Research Infrastructures can benefit from subscription services in several ways. They may set one up to serve their own dedicated end user communities, either by pushing new or updated data sets to their "customers", or by configuring a system that automatically advertises the existence of new data e.g. to those who downloaded previous versions. RIs that require data from other sources, for example to create more elaborated data products, can optimise their internal work flows by signing up to receive automatic updates.

Using VREs to support user-driven data analysis

By providing access to an optimised computational environment which integrates both data and appropriate analysis tools, any data producing organisation can significantly promote the use of its outputs and also help ensure that end users are given the best conditions to perform the analyses and processing that their scientific questions require.

Optimisation of data and metadata interoperability

Funders are pushing strongly for research data (especially publically funded) to be made reusable -- both by other practitioners in the original science domain and also, when possible, by researchers from other fields. This provides a strong argument for RIs to make sure that their data are by default Findable, Accessible, Interoperable and Reusable (FAIR) as part of the normal data management routine. This is indeed a reasonable demand, as many important research discoveries – not the least in the Earth sciences - have been made possible by re-working old data and/or by comparison of old data with recently collected data.

4.3 Individual researchers and others

These include researchers, educators, and the general public – which may have good knowledge of what they want to do, but have access to limited technical and computational resources. VREs and similar platforms will be very important here, as well as well-designed data discovery services. User-friendly data portal designs, supporting easy access to both datasets and related metadata, as well as allowing export in domain-specific formats, are also important.

Near-Real Time data quality control of sensor data

Individual researchers in general perform their own processing, including those associated with quality control and assurance, of sensor data that they collect themselves. However, when accessing data from collaborators they usually rely on QC and QA to have been already applied.

Subscription services for data products

This user category can certainly benefit from signing up to services that automatically advertises the existence of new versions or updates to data that they have downloaded previously – for example annually receiving the observational data from a station for which they have a long-standing scientific interest. Typically, the greatest interest here would be for accessing finalised and aggregated data products created by an RI at its data centre. However, subscription services can also be configured to allow customised processing, bringing an opportunity for research groups to benefit from large-scale cluster computations at the data centre side.

Using VREs to support user-driven data analysis

As outlined in **Chapter 4.2** above, access to an optimised computational environment which integrates both data and appropriate analysis tools, has the potential benefit of ensuring that end users are given the best conditions to perform the analyses and processing that their scientific questions require. This is especially true for VREs that allow individual research groups and researchers that either lack the monetary resources and/or technical knowledge required to set up and maintain high throughput or high performance computation facilities, or have limited access to specific software platforms.

Optimisation of data and metadata interoperability

End users of ENVRIplus data products include both other RIs and environmental and Earth science researchers from across Europe and the rest of the world. Ensuring that all data products, ranging from near real-time data to fully quality controlled finalised datasets, are FAIR is a prerequisite for earning users' trust and instilling their confidence in ENVRIplus.

5 FACILITATING UPTAKE OF ENVRIplus DATA SERVICES & END PRODUCTS

In order to maximize the benefits to both scientific communities and society, it is very important for ENVRIplus service developers and providers to understand the needs and requirements of their potential end user categories. This can be achieved through a "listen, engage, collaborate" strategy, covering service exposure & dissemination through catalogues and discovery services, access to documentation & training, operational support, and the provision of interactive platforms for dialogue and feedback.

Indeed, the WP9 Description of Work requirement to deepen and further develop the contacts between ENVRIplus partner RIs in various domains to their stakeholders and end user communities, in order to jointly map out requirements and to explore solutions to issues relating to data access, replication or transfer. To ensure fruitful and productive interactions, various forms of communication are envisaged, ranging from collaboration on specific use cases, organisation of meetings and workshops, and preparation of documentation and training. Many

of these activities fall under the responsibility of WP9, but also the other Theme 2 work packages are engaged in e.g. use case teams.

In the following, we briefly summarize related activities of WP9, basing the discussion where applicable on the four mechanisms introduced in **Chapter 3**. Importantly, all the end user categories introduced above in **Chapter 4** need to be engaged in a targeted and customised manner, taking into account the differences in technological status, ICT competence and data management experience.

5.1 Engaging stakeholders

5.1.1 Exploring user requirements

Near-Real Time data quality control of sensor data

Stakeholders of numerous environmental RIs have been involved in an initial survey on relevant QC algorithms used in the respective infrastructure. This survey will inform the development of generic implementations for QC algorithms that are of relevance to several infrastructures.

Subscription services for data products

The pilot activity was initiated by the marine research community, which needed an effective yet simple solution for data subscription to serve end users. Requirements included that the service should be easily implemented by any research infrastructure, and that it should be straightforward to integrate with already existing data management practices.

Using VREs to support user-driven data analysis

The Task 9.1 agile teams will help Task 9.2 to reach RIs, e.g., distribute questionnaires on requirements and collect feedback.

Optimisation of data and metadata interoperability

As part of WP 5 activities, questionnaires were sent out in 2015-2016 to all ENVRIplus partners to map out their requirements on the interoperability-related topics of Cataloguing, Semantic information linking, Data identification & citation and Processing. The outcome of this study was reported in deliverable D5.1 [Atkinson 2016]. While some work packages (e.g. WP6) are planning follow-up surveys *with* ENVRIplus partners, there are currently no concrete plans for a comprehensive survey that could map out the corresponding needs and requirements of user communities *outside* of ENVRIplus. However, similar activities are being carried out in the framework of WP9.

5.1.2 Workshops

Near-Real Time data quality control of sensor data

In the short-to-medium term, we aim to primarily engage stakeholders at the twice-yearly ENVRIweek meetings where we plan to present the progress and outcomes of work on this aspect in order to engage partner infrastructures in the ongoing development work. The long-term engagement with other stakeholders, including infrastructures outside of ENVRIplus, remains to be decided, and no dedicated workshops are currently planned.

Subscription services for data products

In order to engage the stakeholders, workshops has been led and demos will be held, e.g. in the framework of the regularly organised ENVRIweek collaboration meetings.

Using VREs to support user-driven data analysis

ENVRIweek meetings provide an important platform for community communication and networking. In the pass ENVRI events, use cases sessions have been organised to present agile implementation results. We will continue to make use of the opportunities to engage the stakeholders. Demonstrations of VRE usability will be presented in the upcoming ENVRIweek.

Agile teams are efficient channels to reach out to RIs. Presentations and talks will be given during community conferences and events. For example, members of the agile team for the IC_3 use case have presented e.g. data processing issues during annual EISCAT_3D user meetings and biannual EISCAT symposia. They will also discuss the technology with related partners and projects.

Optimisation of data and metadata interoperability

In the framework of the November 2016 ENVRIplus Week (Prague, Czech Republic), a workshop was co-organised with Copernicus. The rationale for the meeting was that although many environmental RIs are placing Copernicus on the top-list of their users, very few ENVRIs have clearly formalised specific agreements with Copernicus, directly or indirectly, for any service provision. The workshop provided an opportunity for several RIs across the different domains and at different level of maturity to expose the current status of their interaction with Copernicus. Equally important, representatives from Copernicus services and the EEA were able to outline their expectations on timeliness of data transfer, data quality, and interoperability aspects.

The possibilities to arrange further workshops with Copernicus and other generic initiatives during the remaining ENVRIplus weeks are under investigation. Other options for stakeholder interactions include giving presentations about ENVRIplus data products at e.g. Copernicus collaboration meetings as well as at international and national scientific conferences.

5.1.3 User-driven testing & validation

Near-Real Time data quality control of sensor data

As this service is currently at the proof-of-concept stage, and neither operational nor 'user ready', no mechanism-specific evaluation criteria have been developed.

Subscription services for data products

We will first demonstrate the data flow to push data on a common repository, followed by the integration of Euro-Argo User Interface with the Data Subscription Service. The solution is in progress of integration on a data discovery portal and could serve as a showcase. Once the technical solution has been evaluated, other close stakeholders (SeaDataNet, SOCAT and EMSO) will be invited to join.

Using VREs to support user-driven data analysis

We will track the usability of the pilot results and evaluate the degree of "operationality". From stakeholders, developers and end users, we will collect information on potential user numbers, the Technology Readiness Level (TRL), the resources (computing, storage, staffing, funding) availability for continuing operation, and agreements between service/data/computing/storage providers.

Optimisation of data and metadata interoperability

The FAIR principles should form the basis for defining suitable evaluation criteria related to interoperability. A dedicated FAIRness evaluation and grading system is under development [Doorn 2016], and could be applied also by WP9.

5.2 Disseminating the functionalities of the finalised mechanisms

5.2.1 The ENVRIplus service portfolio

In order to facilitate the discovery of the services the ENVRIplus offers to the ENVRI community and beyond, a Service Portfolio wiki page has been set up – currently hosted at https://confluence.egi.eu/display/EC/ENVRIplus+Service+Portfolios. The portfolio tool allows users to identify services and applications that can be of use to them, based on service name, area and short description, as well as readiness level (phase). The interface will be augmented with links to e.g. longer descriptions, documentation and evaluation scores.

5.2.2 Documentation and training

The responsibility for providing comprehensive documentation on the functionalities of ENVRIplus services rests with the development teams. Examples of materials include summaries of technical specifications, installation instructions, usage manuals, tutorials, and training materials. Importantly, if an instance of a service is deployed by another RI than the original provider, additional instructional materials may be needed, for example outlining local user access control and Authentication, Authorisation and Identification (AII).

In many cases, dedicated training activities – including tutorials, webinars and trainer-led interactive sessions – will be required to facilitate and encourage uptake of the ENVRIplus services. The organisation of such activities, as well as the production of related materials, falls under the responsibility area of Themes 5 ("Knowledge Transfer") and 6 ("Communication & Dissemination").

5.2.3 Workshops and "site visits"

Targeting specific end user communities, a number of meetings or conferences are organized by theme2 or through the coordination with ENVRIplus Themes 5 ("Knowledge Transfer") and 6 ("Communication and Dissemination").

Examples of workshops include:

- Copernicus mini-workshop during the ENVRI week in 2016
- IT4RIs (Interoperable infrastructures for interdisciplinary big data research) workshops (2015 in Munich, 2016 in Porto)
- E-Infrastructure/RI/VRE mini-workshops during the ENVRI weeks in 2016 and 2017 (planned)
- DI4R conference sessions (2016 in Krakow, 2017 in Brussels (planned))

A special category of workshop activities are the so-called ENVRIplus Theme 2 "site visits". These are organised as 2-3 day meetings between researchers and technology experts from an ENVRIplus partner RI, and ICT specialists with expertise on research data management topics (compare **Chapter 1**) with relevance to that RI's current stage of development. During the visits, specific aspects of the RI activities can be discussed, allowing the exchange of experiences and knowledge, as well as the promotion of existing work (service portfolio), the collection of requirements, and opportunities for defining new use cases for consideration by Task 9.1. At the time of writing, site visits have taken place at eLTER, EuroArgo, EPOS and AnaEE, and several more, including ICOS, are being planned.

5.3 Providing support to RIs and users

The Description of Work for Task 9.2 states that "The Task should provide operation support for RIs at both levels of computing and data infrastructure and of RI." We interpret this to mean that WP9 should be prepared to offer

- Information on the functionality, usability and operationality of ENVRIplus services via descriptions at the Service Portfolio web site;
- An interactive web application (coupled to the portfolio) where end users can 1) contribute their own feedback and evaluation information about the services; and 2) explore and view accumulated evaluation scores; and
- Access to further documentation and other support for all ENVRIplus services, including assistance with contacting developers and providers.

6 PLANS FOR THE D9.4 DEMONSTRATOR

The primary objective of Task 9.2 is to track the usability and operational issues of the services deployed by ENVRIplus. In line with this, it is our intention that the final deliverable of Task 9.2, the D9.4 demonstrator, should provide a concrete example of how this can be achieved through collecting relevant validation information and making this available to service developers, service providers and end users.

Various criteria for service evaluation have already been developed in the framework of many other initiatives, for example INDIGO-DataCloud [Indigo 2017], and it would indeed be convenient to base the demonstrator on such pre-existing packages – especially if a majority of the ENVRIplus services are disseminated through e.g. the EGI service portal. However, we need to ensure that the system chosen is able to cover - in sufficient detail – the aspects of usability and operationality that are central to Task 9.2.

Specifically, the implemented evaluation criterion list should as far as possible cover all aspects described in Chapter 5 of the ENVRIplus deliverable D9.1 [Chen 2017], and optimally support also additions aimed at fully covering the task 9.2 usability and operationality aspects. Importantly, the tool should be relevant and useful to both ENVRIplus partners (service developers and providers) and stakeholders (service end users). In addition, the chosen tool should be easy-to-use and allow both to collect, analyse and display relevant information about service evaluations and operational status. Finally, it should ideally be possible to integrate the tool with the ENVRIplus Service Portfolio, for example by linking to dynamically created result summaries or visualisations of evaluation score statistics.

We therefore plan to start by making an inventory of existing service evaluation tools, with special focus on user friendliness, aspects covered and possibilities for customising both the user interface and the output formats. If no suitable alternative is identified, we are prepared to develop our own basic, web-based application that can support the needs of WP 9 (i.e., both Tasks 9.1 and 9.2) with respect to service evaluation and validation.

7 CONCLUSIONS

The work in WP9 is essential to evaluate the uptake of the activities in ENVRIplus into the environmental research infrastructures and their user communities. A large number of these activities are related to Theme 2, including all the services embedded, further developed and tested in the use and implementation cases of Work Packages 5-9.

Indeed, at first sight, the long list of services and use cases under study can be quite bewildering. However, once we associate the cases with different research data management topics, as well as with phases of the research data lifecycle, different kinds of patterns start to emerge. Further filters can be applied, for example on the end user categories involved, clarifying the picture even further. In this report, we distinguish four generic "mechanisms" of data sharing: 1) Near-Real Time data quality control of sensor data; 2) Subscription services for data products; 3) Using VREs to support user-driven data analysis; and 4) Optimisation of data and metadata interoperability.

Based on the descriptions of related use cases, their current level of development, as well as the experiences of the associated agile teams, we have sought to draw conclusions on a number of aspects with major importance for the two overarching goals of WP9: the validation of the services themselves, and the evaluation of their uptake by the scientific community. These aspects include 1) who the most important stakeholders and end user communities are; 2) how to best engage with them; and 3) how to effectively assess usability and "operationality" of the services. The outcome of this analysis is now informing our follow-up actions, including the design and implementation of D9.4 demonstrator, the interactive web tool for service evaluation.

The design, development and deployment of services by the use cases involved in the various mechanisms is still on-going, and it is still too early to confidently estimate the degree of technical readiness level that the resulting services will be able to achieve. Nevertheless, we feel confident that end users of all categories will greatly benefit from the inclusion of comprehensive information on both validation and evaluation activities into the Service Portfolio. The demonstrator for service evaluation will be an important component supporting the uptake of the ENVRIplus data services and products as highlighted in the different use and implementation cases.

8 IMPACT ON PROJECT

This report provides an analysis of how Theme2 service results can be viewed from the point of view of RIs. It identifies four realistic mechanisms, or usage scenarios, in which end user communities of different types and sizes could benefit from those services. These can help RIs better understand the concepts and usage of the services, and make it easy for them to adopt them. The report also develops strategies for community engagements. It identifies targeting user groups, establishes dissemination channels for reaching out to the different user

communities, and defines a mechanism for tracking the usefulness of those services. These efforts help to explore the full potential of Theme 2 service results, and also promote the development results to the community.

9 IMPACT ON STAKEHOLDERS

This deliverable targets representatives of a wide range of end users of ENVRIplus services, ranging from large, "generic" organisations like GEOSS and Copernicus, through research infrastructures (including the ENVRIplus RIs) to research groups. We expect RIs to acknowledge and get involved in the dissemination activities coordinated by T9.2. RIs should actively try the products & services developed by ENVRIplus and 1) provide feedback and validation information; 2) take into consideration the adoption of relevant solutions into their daily practices; and 3) explore additional usage, also outside of the originally intended research domains. RIs should also help to reach out to any connected projects and networks -- to maximise the impacts of the ENVRIplus project outcomes, efforts and support from the whole end user community is needed.

REFERENCES

[Atkinson 2016] M. Atkinson, A. Hardisty, R. Filgueira, C. Alexandru, A. Vermeulen, K. Jeffery, T. Loubrieu, L. Candela, B. Magagna, P. Martin, Y. Chen and M. Hellström: A consistent characterisation of existing and planned RIs. ENVRIplus Deliverable 5.1, submitted on April 30, 2016. Available at http://www.envriplus.eu/wp-content/uploads/2016/06/A-consistent-characterisation-of-RIs.pdf

[Bechhofer 2013] S. Bechhofer, I. Buchan, D. De Roure, P. Missier et al.:. Why linked data is not enough for scientists. Future Generation Computer Systems, Volume 29, Issue 2, February 2013, Pages 599-611, ISSN 0167-739X, http://dx.doi.org/10.1016/j.future.2011.08.004.

[Candela 2017] L. Candela, G. Coro, P. Pagano, G. Panichi, M. Atkinson, R. Filgueira, D. Bailo, C-F. Enell, M. Fiebig, F. Haslinger, M. Hellström, A. Vermeulen, H. Lankreijer, R. Huber, S. Joussaume, F. Guglielmo and V. Mendez: Interoperable data processing for environmental RI projects: system design. ENVRIplus deliverable D7.1, submitted on March 27, 2017. Available at http://www.envriplus.eu/wp-content/uploads/2015/08/D-7.1.pdf.

[Chen 2017] Y. Chen, B. Grenier, M. Hellström, A. Vermeulen, M. Stocker, R. Huber, B. Magagna, I. Häggström, M. Fiebig, P. Martin, D. Vitale, G. Judeau, T. Carval, T. Loubrieu, A. Nieva, K. Jeffery, L. Candela and J. Heikkinen: Service deployment in computing and internal e-Infrastructures. ENVRIplus Deliverable 9.1, submitted on August 31, 2017. Available at http://www.envriplus.eu/wp-content/uploads/2015/08/D9.1-Service-deployment-in-computing-and-internal-e-Infrastructures.pdf.

[Doorn 2016] P. Doorn and I. Dillo: FAIR Data in Trustworthy Data Repositories. Webinar co-organised by DANS, EUDAT and OpenAire, December 13, 2016. Recording and slides available at https://eudat.eu/events/webinar/fair-data-in-trustworthy-data-repositories-webinar. Accessed 2017-10-26.

[ENVRI RM V2.1 2016] ENVRI Reference Model V2.1, November 9 2016. https://wiki.envri.eu/download/attachments/8553250/EC-091116-1403.pdf. Accessed 2017-01-10. Also available in wiki format at https://wiki.envri.eu/display/EC/ENVRI+Reference+Model.

[ENVRIplus 2015a] ENVRIplus Description of Work (DoW), public part. ENVRIplus Grant Agreement, Annex 1, part A. Horizon 2020 project no. 654182. Associated with document Ref. Ares(2015)1488547. Available at http://www.envriplus.eu/wp-content/uploads/2015/08/ENVRIplus_DoW_public.pdf.

[ENVRIplus 2015b] ENVRIplus project description, public part. ENVRIplus Grant Agreement, Annex 1, part B. Horizon 2020 project no. 654182. Associated with document Ref. Ares(2015)1488547. Available at http://www.envriplus.eu/wp-content/uploads/2015/08/ENVRIplus_PartB_public.pdf.

[FORCE11 2014] FORCE11 project: Guiding Principles for Findable, Accessible, Interoperable and Re-Usable Data Publishing Version B1.0. https://www.force11.org/fairprinciples. Accessed 2016-11-28.

[Hellström 2016] M. Hellström, A. Vermeulen, O. Mirzov, S. Sabbatini, D. Vitale, D. Papale, J. Tarniewicz, L. Hazan, L. Rivier, S.D. Jones, B. Pfeil and T. Johannessen: Near Real Time Data Processing In ICOS RI. In Proceedings of 2nd International Workshop on Interoperable infrastructures for interdisciplinary big data sciences (IT4RIs 16), Porto, Portugal November 30, 2016. Zenodo. http://doi.org/10.5281/zenodo.204817

[Hellström 2017] M. Hellström, M. Lassi, A. Vermeulen, R. Huber, M. Stocker, F. Toussaint, M. Atkinson and M. Fiebig: A system design for data identifier and citation services for environmental RIs projects to prepare an ENVRIplus strategy to negotiate with external organisations. ENVRIplus deliverable D6.1, submitted on January 31, 2017. Available at http://www.envriplus.eu/wp-content/uploads/2015/08/D6.1-A-system-design-for-data-identifier-and-citation-services-for-environmental-RIs.pdf.

[Huber 2017] R. Huber, O. Gilbert, A. Vermeulen, D. Papale, J.-F. Rolin, C. Waldmann and S. Rohs: Report for best practices on robust telecom/data transmission. ENVRIplus deliverable D3.3. Submitted on April 30, 2017. Available at http://www.envriplus.eu/wp-content/uploads/2015/08/D3.3.pdf.

[INDIGO 2017] Members of the INDIGO-DataCloud collaboration: Exploitation analysis based on agreements made and usage statistics. Deliverable D2.9 of INDIGO-DataCloud RIA-653549, dated October 4, 2017. Available at https://owncloud.indigo3datacloud.eu/index.php/s/LJuCHX54yUZM6UL.

[INSPIRE 2017] Infrastructure for Spatial Information in Europe (INSPIRE). The INSPIRE knowledge base: About INSPIRE. https://inspire.ec.europa.eu/about-inspire/563. Accessed on 2017-10-23.

[Jeffrey 2017] K.G Jeffery, Z. Zhao, B. Magagna, A. Nieva de la Hidalga, L. Candela, C.-F. Enell, M. Hellström, A. Hardisty, C. Paxton and F. Toussaint: Data Curation in System Level Sciences: Initial Design. ENVRIplus deliverable D8.1. Submitted on January 31, 2017. Available at http://www.envriplus.eu/wp-content/uploads/2015/08/D8.1-Data-Curation-in-System-Level-Sciences-Initial-Design.pdf.

[Martin 2015] P. Martin, P. Grosso, B. Magagna, H. Schentz, Y. Chen, A. Hardisty, W. Los, K. Jeffrey, C. de Laat, Z. Zhao: Open Information Linking for Environmental Research Infrastructures. Paper presented at the First International workshop on Interoperable infrastructures for interdisciplinary big data research (IT4RIs). Proceedings of IEEE eScience, Munich Germany, p546-553 , 2015. http://dx.doi.org/10.1109/eScience.2015.66.

[Martin 2017] P. Martin, Y. Chen, A. Hardisty, K. Jeffery, and **Z.** Zhao, *Computational Challenges in Global Environmental Research Infrastructure,* in Terrestrial Ecosystem Research Infrastructures: Challenges, New developments and Perspectives, A. Chabbi and H.W. Loescher, Eds. CRC Press, 2017,isbn 9781498751315. See https://www.crcpress.com/Terrestrial-Ecosystem-Research-Infrastructures-Challenges-and-Opportunities/Chabbi-Loescher/p/book/9781498751315.

[Petzold 2015] A. Petzold, V. Thouret, C. Gerbig, A. Zahn, C.A.M. Brenninkmeijer, M. Gallagher, M. Hermann, M. Pontaud, H. Ziereis, D. Boulanger, J. Marshall, P. Nédélec, H.G.J. Smit, U. Friess, J.-M. Flaud, A. Wahner, J.-P. Cammas, A. Volz-Thomas & IAGOS TEAM: Global-scale atmosphere monitoring by in-service aircraft – current achievements and future prospects of the European Research Infrastructure IAGOS. Tellus B: Chemical and Physical Meteorology Vol. 68 , Issue s1, 2016. http://dx.doi.org/10.3402/tellusb.v67.28452@zelb20.2016.68.issue-s1.

[Rönkkö 2015] M. Rönkkö, O. Kauhanen, M. Stocker, H. Hytönen, V. Kotovirta, E. Juuso, and M. Kolehmainen: Quality Control of Environmental Measurement Data with Quality Flagging. Environmental Software Systems. Infrastructures, Services and Applications, IFIP Advances in Information and Communication Technology 448, pp. 343-350. https://doi.org/10.1007/978-3-319-15994-2_34

[Koulouzis 2017] S. Koulouzis, P. Martin, T. Carval, B. Grenier, G. Judeau, J. Wang, H. Zhou, C. de Laat, and Z. Zhao: Seamless Infrastructure Customisation and Performance Optimisation for Time-critical Services in Data Infrastructures. To be published in the proceedings of the Eighth International Workshop on Data-Intensive Computing in the Clouds, Denver, CO, November 12, 2017. In cooperation with ACM SIGHPC 2017.

[Vejen 2002] F. Vejen (ed), C. Jacobsson, U. Fredriksson, M. Moe, L. Andresen, E. Hellsten, P. Rissanen, T. Palsdottir, and T. Arason: Quality Control of Meteorological Observations. Automatic Methods Used in the Nordic Countries. Climate Report 8/2002, Norwegian Meteorological Institute, 2002.

[Wilkinson 2016] M. D. Wilkinson, M. Dumontier, I.J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J-W. Boiten, L. B. da Silva Santos, P.E. Bourne, J. Bouwman, A.J. Brookes, T. Clark, M. Crosas, I. Dillo, O. Dumon, S. Edmunds, C.T. Evelo, R. Finkers, A. Gonzalez-Beltran, A.J.G. Gray, P. Groth, C. Goble, J.S. Grethe, J. Heringa, P.A.C. 't Hoen, R. Hooft, T. Kuhn, R. Kok, J. Kok, S.J. Lusher, M.E. Martone, A. Mons, A.L. Packer, B. Persson, P. Rocca-Serra, M. Roos, R. van Schaik, S-A. Sansone, E. Schultes, T. Sengstag, T. Slater, G. Strawn, M.A. Swertz, M. Thompson, J. van der Lei, E. van Mulligen, J. Velterop, A. Waagmeester, P. Wittenburg, K. Wolstencroft, J. Zhao and B. Mons: The FAIR Guiding Principles for scientific data management and stewardship, Sci. Data 3:160018 doi: http://dx.doi.org/10.1038/sdata.2016.18.

[Wong 2015] A. Wong, R. Keeley, T. Carval and the Argo Data Management Team. Argo Quality Control Manual for CTD and Trajectory Data. http://dx.doi.org/10.13155/33951

[Zhao 2015] Z. Zhao: The theme of data for science. Presentation at the 1st ENVRIPLUS week meeting, Prague, Czech Republic, November 16-20, 2015.

APPENDIX A. ACRONYMS & TERMS USED IN THIS REPORT

AAI: Authentication, Authorisation and Identification.

AnaEE: Analysis and Experimentation on Ecosystems – an ENVRIplus partner. See http://www.anaee.com/.

API: Application Programming Interface.

B2x: Refers to the B2 service suite (including **B2FIND**, **B2SHARE**, **B2SAFE**) developed under the EUDAT project, see https://eudat.eu/services-support/.

C3S: Copernicus Climate Change Service, http://climate.copernicus.eu/.

CAMS: Copernicus Atmospheric Monitoring Services, http://copernicus.eu/main/atmosphere-monitoring/.

CDI: Collaborative Data Infrastructure, the "successor" of the EUDAT project.

CODATA: The Committee on Data for Science and Technology, see http://www.codata.org/.

Copernicus: Previously known as GMES (Global Monitoring for Environment and Security), this is the European Programme for the establishment of a European capacity for Earth Observation, see http://www.copernicus.eu/.

CSC: The Finnish national IT Center for Science, https://www.csc.fi/.

D4Science: An e-Infrastructure offering Hybrid Data Infrastructure service and Virtual Research Environments. http://www.d4science.org/.

DataONE: Data Observation Network for Earth, see https://www.dataone.org/.

DIRAC: Distributed Infrastructure with Remote Agent Control, a workload management service for handling large volumes of data, see http://diracgrid.org/.

DOI: Digital Object Identifier.

DRIP: Dynamic real-time infrastructure planner. A infrastructure optimisation service developed in WP7.

EC: Eddy Covariance, a measurement technique for obtaining exchange fluxes of e.g. greenhouse gases and energy between the Earth surface and the atmosphere.

ECMWF: European Centre for Medium-Range Weather Forecasts. https://www.ecmwf.int/.

EEA: The European Environmental Agency, https://www.eea.europa.eu/.

EGI: EGI.eu – an ENVRIplus partner, http://egi.eu/.

EISCAT: European Incoherent Scatter Scientific Association - an ENVRIplus partner, see https://www.eiscat.se/.

EISCAT_3D: the next generation incoherent scatter radar now being built by EISCAT, see https://www.eiscat3d.se.

eLTER: Long-Term Ecosystem Research in Europe – an ENVRIplus partner RI. See http://www.lter-europe.net/.

EMSO: European Multidisciplinary Seafloor and water-column Observatory, http://www.emso-eu.org/.

EOSC: European Open Science Cloud, an initiative from the European Commission see https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud.

EOSC-Hub: a H2020 EINFRA12(A) project, contributing to the European Open Science Cloud implementation. https://www.egi.eu/about/newsletters/introducing-the-eosc-hub-project/.

ePIC: The European Persistent Identifier Consortium, a PID service provider using the Handle system, http://www.pidconsortium.eu/.

EPOS: European Plate Observing System – an ENVRIplus partner RI. See https://www.epos-ip.org/.

EUDAT: A Horizon2020 project and ENVRIplus partner, https://eudat.eu/.

EuroARGO: European contribution to the Argo programme – an ENVRIplus partner. See http://www.euro-argo.eu/.

EuroGOOS: European Global Ocean Observing System – an ENVRIplus partner. See http://eurogoos.eu/.

ETL: Extract, Transform, Load - a process in database usage, especially in data warehousing. See e.g. https://en.wikipedia.org/wiki/Extract,_transform,_load.

FAIR: Findable, Accessible, Interoperable and Reusable.

Fluxnet: A global network of micrometeorological tower sites that use eddy covariance methods to measure the exchanges of carbon dioxide, water vapour, and energy between terrestrial ecosystems and the atmosphere. See https://daac.ornl.gov/cgi-bin/dataset_lister.pl?p=9.

FORCE11: FORCE11 is a community of scholars, librarians, archivists, publishers and research funders that works toward improved knowledge creation and sharing. See https://www.force11.org.

GBIF: Global Biodiversity Information Facility, see https://www.gbif.org/.

GEOSS: Global Earth Observation System of System, see https://www.earthobservations.org/geoss.php.

GHG: Greenhouse Gas, a gas like CO_2, CH_4 and N_2O that when present in the atmosphere contributes to the global greenhouse effect.

GUI: Graphical User Interface.

Handle: A system for registering persistent identifiers, see http://handle.net/.

HPC: High Performance Computing.

HTTP: Hyper Text Transfer Protocol.

IAGOS: In-Service Aircraft for a Global Observing System – an ENVRIplus partner. See https://www.iagos.org/.

ICOS: Integrated Carbon Observation System – an ENVRIplus partner. See https://icos-rii.eu/.

ICSU-WDS: The International Council for Science's World Data System, see https://www.icsu.org.

ICT: Information and Communications Technology.

INDIGO: INtegrating Distributed data Infrastructures for Global ExplOitation. INDIGO-DataCloud is a Horizon2020 project, see https://www.indigo-datacloud.eu/.

iRODS: Integrated Rule-Oriented Data System, an Open Source Data Management Software, see https://irods.org/.

JSON: JavaScript Object Notation.

NeIC: Nordic e-Infrastructure Collaboration. https://neic.no/.

NOAA: The United States' National Oceanic and Atmospheric Administration, see http://www.noaa.gov/.

NREN: National Research and Education Network. https://www.geant.org/About/NRENs.

NRT: Near Real-Time.

OIL-E: Open Information Linking for Environmental research infrastructures, an ontology representation of the ENVRI Reference Model and mappings between RM and other domain specific ontologies, http://oil-e.net/ontology/.

PaaS: Platform as a Service. https://en.wikipedia.org/wiki/Platform_as_a_service.

PID: Persistent Identifier. https://en.wikipedia.org/wiki/Persistent_identifier.

QA: Quality Assurance.

QC: Quality Control.

QUARTOD: A project operated by NOOA, see https://ioos.noaa.gov/project/qartod/.

OGC: Open Geospatial Consortium, http://www.opengeospatial.org/.

RDA: The Research Data Alliance, see http://rd-alliance.org/.

RDM: Research Data Management.

RI: Research Infrastructure.
SensorML: Sensor Modelling Language.
SMOS: Soil Moisture and Ocean Salinity.
SOCAT: Surface Ocean CO_2 Atlas, see https://www.socat.info/.
SOS: Sensor Observation Service.
SSN: Semantic Sensor Network.
SSNO: Semantic Sensor Network Ontology.
SWE: Sensor Web Enablement.
TRL: Technology Readiness Level
T-SOS: Transactional Sensor Observation Service.
VRE: Virtual Research Environment.
WPS: Web Processing Service.
XML: Extensible Markup Language.

APPENDIX B. NEAR-REAL TIME DATA QUALITY CONTROL (MECHANISM 1)

Contributed by Markus Stocker and Robert Huber

Summary

Motivation

Since data quality is a key factor of data fitness for use by research communities, ensuring high data quality is a critical data curation activity in any RI. This aspect develops a generic software architecture for near-real time (NRT) quality control (QC) and proposes and implementation of the architecture as a concrete demonstrator. Key features of the architecture are the standardization of observational data and the generic implementation of QC routines that can be reused across RIs. Hence, this aspect explores the possibility of NRT QC-as-a-Service that infrastructures can flexibly configure to suit their observational data and quality control needs. Key requirement is the reuse of QC routine implementations across infrastructures. This is enabled by standardizing the representation of observational data prior to processing such data for QC. Such standardization is implemented through representation translation components that convert incoming observational data in heterogeneous format into a standardised representation for QC processing as well as convert quality annotated processed observational data into output data in heterogeneous format. NRT processing is enabled by a message queue and a distributed real-time computation system.

Technologies involved

The architecture is independent of concrete technologies. However, the proposed implementation adopts a number of concrete technologies. First, EGI serves as the e-Infrastructure that executes NRT QC-as-a-Service. EGI provides required hardware and software components. Regarding software components, EGI provides access to virtual machines, the Apache Storm cluster configured on virtual machines, as well as the messaging service. Apache Storm[7] serves as the implementation of the distributed real-time computation system. Complex NRT QC processing is implemented as Apache Storm Topologies - i.e. a directed graph of processing units, specifically Storm Bolts, each implementing a QC routine. Storm Bolts are implemented in Java. Furthermore, translation components are needed to convert observational data between representations. These are implemented on the EGI messaging service, potentially in various forms (e.g. in Java or as scripts).

Goals related to improving

Since our current approach to implement a standards-based generic NRT QC service is a proof of concept and not a production-ready service, it has to be improved to meet community expectations. In particular, the service requires in-depth validation within real-world QC workflows of one or more RIs, ideally benchmarking with existing domain procedures. In addition, several technical improvements are required before the service can be exploited within RI QC routines. Most importantly, the QC result output needs to be adapted to the requirements of individual RIs. As there is no common quality flagging system, domain-specific QC flag configuration options should be implemented within the service. Currently flagged T-SOS XML,

[7] http://storm.apache.org

the output format needs more thought and development. A simple JSON format resembling the SSN ontology is planned. The number of supported input formats could also be improved to support, e.g., domain specific formats. Most importantly, to improve the overall usability of the service, several user clients and respective GUIs are required. Specifically, a user frontend that supports registering NRT streams and associated sensor specifications is needed. Finally, our service would benefit from a visualisation and plotting utility through which NRT QC flagging could be followed in real time.

Data & metadata access for human users

This aspect doesn't seem to improve data or metadata *access* for either human users or computer-driven processes. It does perhaps improve on metadata quality through an additional annotation representing the quality of the data. This annotation is both readable by human users as well as in computer-driven processes.

Data & metadata access for computer-driven processing

This aspect doesn't seem to improve data or metadata *access* for either human users or computer-driven processes. It does perhaps improve on metadata quality through an additional annotation representing the quality of the data. This annotation is both readable by human users as well as in computer-driven processes.

Context:

Couplings to other ENVRIplus activities

This aspect relates to Theme 1, specifically WP 1, WP 2, and WP 3. The aspect is committed to standardizing observational data streams, ideally at the sensor level so that devices directly generate data following one or more suitable standard format (e.g. OGC SWE). Since this aspect deals with observational data processing, it furthermore relates to Theme 2, in particular WP 7. Finally, the aspect relates to TC 4 on Sensor Registry. Such registries can provide data about sensors, such as sensor capabilities, which are important to configure QC routines, e.g. for range checks.

Similar projects outside of ENVRIplus?

Whereas adoption and implementation of OGC SWE or SSN ontology based data transfer and management frameworks become increasingly important for environmental RIs, our approach to use these standards to ease NRT QC is novel. However, some initiatives such as QARTOD[8] are currently trying to harmonise NRT QC routines within them marine domain and RIs such as ICOS [Hellström 2016], EuroARGO [Wong 2015], EMSO, EuroGOOS[9] or IAGOS [Petzold 2015]. These initiatives are an important prerequisite for the identification and provision of common and generic NRT routines but no cross-disciplinary approach to solve the issue of NRT QC exists so far. No attempt to implement technical solutions or generic services has been initiated previously.

[8] https://ioos.noaa.gov/project/qartod/

[9] http://eurogoos.eu/data-management-exchange-quality-working-group-data-meq/

Importance

Outside of the current domain (atmosphere, ecosystems, marine, solid earth)

Clearly, QC - possibly in a NRT context, e.g. for early warning or situation awareness systems - on observational or other kinds of data is critical in most domains. Checking for quality increases the fitness of data in downstream use, for instance by research communities. However, as the current focus is on environmental RIs, the schema used for standardised data is tailored for observational data, specifically data generated by sensors and, primarily, sensing devices (although easily applicable for human observers). While observations are obviously made in other scientific and research contexts, for instance in the social sciences via surveys and interviews, the current schema would probably have to be adapted to suit the needs of the different context. In contrast to the environmental and earth sciences, other domains may have less of a requirement for NRT processing and rely more on batch processing for QC. The different mode would arguably require a different architecture or at least a different set of technologies. Specifically, in batch processing mode we would probably not include a distributed real-time computation system such as Apache Storm. Rather, the system would surely rely on more classical ETL processing on database systems.

Regionally (Europe) and Globally

It is unclear whether there is a difference between regional and global significance, as the importance of the problem and possible approaches are independent of region. NRT QC is a common theme in RIs, globally. In conjunction with responses to our survey on specific practices among ENVRIplus partners, we have received many critical comments regarding the feasibility of generic NRT QC services. Serious concerns exist among domain specialists. Tacit knowledge is of particular concern as too often information relevant to the parametrization of QC algorithms depends on experience that needs to first be acquired and may not be documented. However, as seen in **Table 3** our survey highlights communalities in the QC algorithms used among different infrastructures. (In the table, green is used to highlight those algorithms that are used by a majority of the RIs, as evidenced by the number indicated in the rightmost "#" column.)

Commonalities are also a key driving factor behind proposed global standardizations of QC patterns, such as those proposed by QARTOD. For numerous kinds of QC tests - e.g. gap test, impossible location, spike test or rate of change test - QARTOD has documented generic patterns that are independent of specific implementations (e.g. programming language) as well as context (though patterns may require contextual parameterization). For instance, the pattern for Gap Test is as follows:

> *Gap Test:*
> *Def : NOW - TIM_STMP > TIM_INC , flag=4*
> *where: TIM_INC = Test specifications to be established locally by operator*

Its definition relies on current time (as obtained by some system, possibly the sensing device itself) and subtracts the latest time stamp for which an observation was made (possibly retained in sensing device memory). This difference should not exceed a parameterised time increment. As we can see, this time increment needs to be established by an operator based on context (most importantly device operating frequency). Regardless of contextual parameterization, such standardised patterns for QC tests are useful here since they provide a baseline for generic QC routine implementations.

TABLE 3. COMMON NRT QC ROUTINES USED BY ENVRI RESEARCH INFRASTRUCTURE PARTNERS [HUBER 2017]. GREEN ROWS HIGHLIGHT THE MOST COMMON TEST ALGORITHM TYPES, USED BY 5 OR MORE RIS.

	EMSO	FixO3	EuroArgo	ACTRIS (surf.)	ACTRIS (lidar)	IAGOS	ICOS (ETC)	ICOS (OTC)	GROOM	Frequency of use
Data Integrity Test	-	-	-	x	-	-	x	-	-	2
Metadata Consistency Test	-	-	-	x	x	x	x	-	-	4
Platform/Sensor Identification	x	x	x	x	-	-	x	-	x	6
Date/Time Check	x	x	x	-	-	x	x	x	x	7
Location Check	-	x	x	-	-	x	-	x	x	5
Spike or Outlier Test	x	x	x	x	-	x	x	x	x	8
Gradient Test	x	x	x	-	-	-	-	-	x	4
Stuck or Constant Value Test	-	x	x	-	-	x	x	x	x	6
Digit Rollover Test	-	-	x	-	-	-	-	-	x	2
Gap Test	-	-	-	-	-	x	x	-	-	2
Rate of Change / Step Test	-	x	-	-	-	x	x	-	-	3
Range Test (Regionality/Past Values)	-	x	x	-	-	x	-	x	x	5
Range Test (Global)	-	x	x	-	-	x	x	-	-	4
Range Test (Instrument Limits)	-	-	-	x	-	x	x	x	-	4
Range Test (Implicit)	-	-	-	x	x	x	-	-	-	3

Discussing a quality flagging scheme, [Rönkkö 2015] highlighted that QC performed prior to data storage can be automated while subsequent quality checks are performed off-line and may require human interaction. Rönkkö et al. adopt the quality flagging scheme by the Nordic meteorological institutes [Vejen 2002] which builds on four QC levels, namely Q0, Q1, Q2, and HQC. Q0 is performed by the sensor or station and Q1 by the data acquisition system prior to data storage. These two quality control levels are performed in real-time and are automated. QC2 is performed by the data management system and HQC by human operators. Both QC2 and HQC are performed off-line. The aspect discussed here focuses on the automated Q0 and Q1 QC levels performed in NRT and suggests the possibility for e-Infrastructures to provide such QC as a flexibly configurable generic service.

Target group

Stakeholders

Currently, the primary stakeholders are surely environmental RIs, specifically those that collect observational data from sensors, sensor networks, monitoring stations. Secondary stakeholders are environmental RIs, possibly virtual infrastructures, which acquire data from other environmental RIs, especially those that acquire data in NRT. RIs in other domains, e.g. life sciences or high-energy physics, that acquire data in real-time are possible stakeholders. Massive observational data are acquired in gene sequencing or in high-energy physics, and such data undergo NRT QC either at the instrument level or in the data acquisition phase following data

collection. In addition to infrastructures that acquire observational data through in-situ monitoring, stakeholders relevant to this aspect also include infrastructures that acquire observational data through remote sensing, specifically satellite based systems for Earth Observation. Such observational data undergo QC before downstream use.

While the focus of this aspect is primarily on observational data, the proposed architecture is not limited to such kind of data and can be applied to other kinds of data acquired in NRT, in particular computational data generated, e.g., in simulations. Computational data need to undergo QC to check data generating software for errors. Since QC is a fundamental task in data acquisition and curation that contributes substantially to increasing data fitness in downstream use, all stakeholders that utilise resulting quality controlled data benefit from upstream QC. Relevant stakeholders include research communities as well as governmental agencies and other authorities, and more broadly user communities, including citizen scientists - at least to the extent that these users benefit from higher quality data.

Plans for engaging the stakeholders

Questionnaires on requirements

Stakeholders of numerous environmental RIs have been involved in an initial survey on relevant QC algorithms used in the respective infrastructure. This survey will inform the development of generic implementations for QC algorithms that are of relevance to several infrastructures.

Workshops and other meetings

In the short-to-medium term, we primarily engage stakeholders through ENVRIplus at the twice-yearly ENVRIweek meetings where we plan to present the progress and outcomes of work on this aspect in order to engage partner infrastructures. However, we must first develop the proof of concept to a maturity level where it can be presented as a credible approach. Hence, we have so far not made concrete plans for stakeholder engagement. Even more unclear is the long-term, and the inclusion of infrastructures, beyond ENVRIplus.

Criteria for evaluation of usability and degree of “operationality”

This does not apply since the service is a proof of concept and not yet operational nor 'user ready'. Required improvements are listed above.

APPENDIX C: DATA SUBSCRIPTION SERVICE (MECHANISM 2)

Contributed by Glenn Judeau, Thierry Carval, Jani Heikkine and Chris Ariyo, with additional material from Spiros Koulouzis, Paul Martin and Zhiming Zhao

Summary

Motivation

Many research infrastructures aim for providing easy to use and efficient data distribution. Researchers need this feature to collect data for their analyses, models, and studies. Due to accumulating data sources, end-users frequently repeat selection and download of the data they need depending on their criteria to keep their collection up to date. There is a need to avoid this recurring and time consuming task.

Moreover, environmental data lifecycle is complex and reflects various stages needed to produce, acquire, cure, publish, and process environmental data. Therefore, the challenging point of this aspect is to enable delivery of frequently updated data to end-users: research infrastructures, scientific communities, and researchers.

Technologies involved

To meet these challenges, the proposed architecture combines various technologies to interconnect the stakeholders through a subscription model, see **Figure 3** below.

It is assumed that data is registered in a repository (B2SAFE) supporting well-defined data identification mechanism. In the proposed architecture, the repository provides an HTTP API for data ingestion and data is identified through Handle/EPIC PIDs [Hellström 2017].

Consequently, end-users could access and subscribe to the data through the use of a web portal provided by the research infrastructure or another data discovery interface. Further, the web portal uses an HTTP API to create and manage subscriptions.

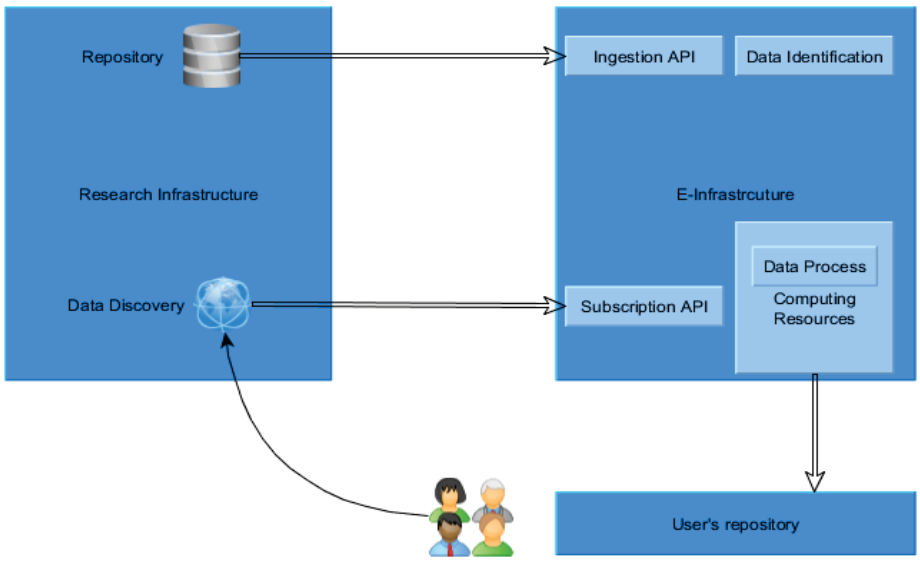

FIGURE 3. THE CONCEPT OF THE DATA SUBSCRIPTION SYSTEM, DESCRIBED FROM THE POINT OF VIEW OF THE END USER..

A data delivery process then involves clustered computing resources to extract data of interest and to produce results that are pushed on to user's cloud repository (B2DROP). Once result datasets are provided, the user receives a notification (by mail in a first step). Moreover, implementation of a citation scheme attached to the delivered data will be investigated.

While the subscription model inherently decouples researchers from monitoring effects of data changes, the response time from the moment of activating a subscription to accessing updates depends on a number of factors including the network data transfer latencies, provisioning time of processing tasks and required resources, the processing itself, and error handling and recovery.

Using the infrastructure optimization component DRIP, developed in WP7, a prototype [Koulouzis et al., 2017] has been made to demonstrate how the production of data can be scheduled using a well-selected set of virtual machines from EGI, and with different ordering of the subscriptions to meet the possible time constraints given by the users or workflows – see **Figure 4**.

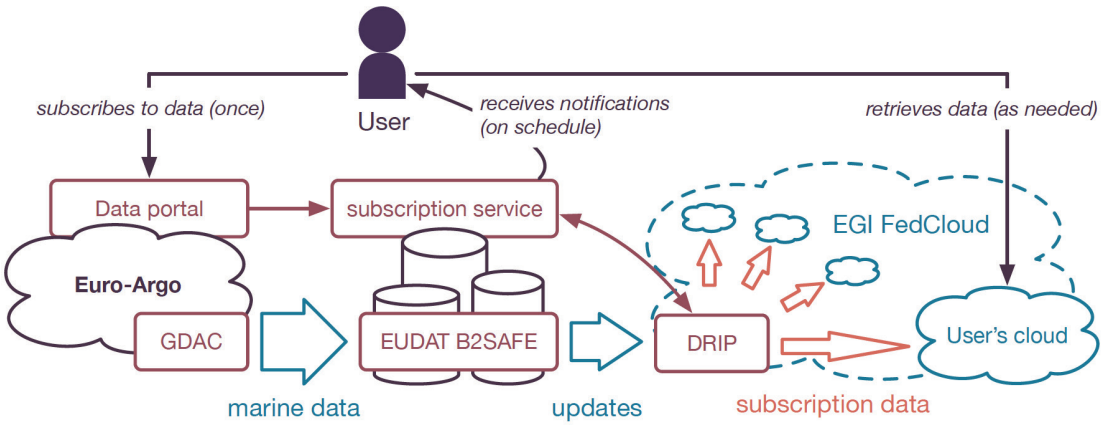

FIGURE 4. THE ARCHITECTURE OF A PROTOTYPED DEMONSTRATOR [KOULOUZIS 2017].

Goals related to improving

Data extraction from a common repository is performed using clustered computation resources. Both HTTP APIs are developed on e-infrastructure side, one for pushing data, the other to manage data subscriptions. Metadata from data collection are discoverable through a web portal.

Global solution is still a proof of concept and next step will be the integration of web portal to the subscription service.

Data & metadata access for human users

The aim of this aspect is to minimise/reduce time consuming data monitoring process for researchers. Initial experiments did with T7.2 show that scheduling the order of subscriptions based on the priority, specific data attributes in the subscriptions, and their size can improve the performance of data product generation for subscriptions [Koulouzis, et al, 2017].

Data & metadata access for computer-driven processing

Such subscription service could be addressed by computer-driven processes (in order to keep their data collection up to date). This could lead to a new way to keep models and analyzes consistent with frequent updates.

Context:

Couplings to other ENVRIplus activities

- Data and metadata interoperability: the more data and metadata are interoperable, the most data source origins could benefit from this service. By gathered data to common resources we can provide services that could lead to new trans-disciplinary analyze, models, and studies.
- Data identification: the more common data identification methods are used, the more data producers and data consumers could benefit from this service.
- VRE support for user driven analysis : Storing data collections in a convenient space that could be associated with storage and CPU can provide advanced services.

Importance

This aspect could lead to a common input stream or digital collaborative space for researchers and data providers.

Target group

Stakeholders

The aspect focuses on Euro Argo and Copernicus Marine Service for the use-case. The objective is to involve SeaDataNet, SOCAT and EMSO in a next step.

Standardisation of input data files could lead to any data producer who wants to diffuse and offer a subscription service to their users. Many research communities have their own data selection portal. We can propose them an easy to embed solution between their portals and digital workspaces, mutualizing resources and leading to a transdisciplinary environment as shown in **Figure 5** below.

Other targeted domains could include:

- Environmental monitoring and forecasting: EU ocean-atmosphere models.
- Calibration and validation of observations: SMOS, Sentinel3 satellite missions.
- Trans-disciplinary communities (ocean, atmosphere, biology, solid earth).

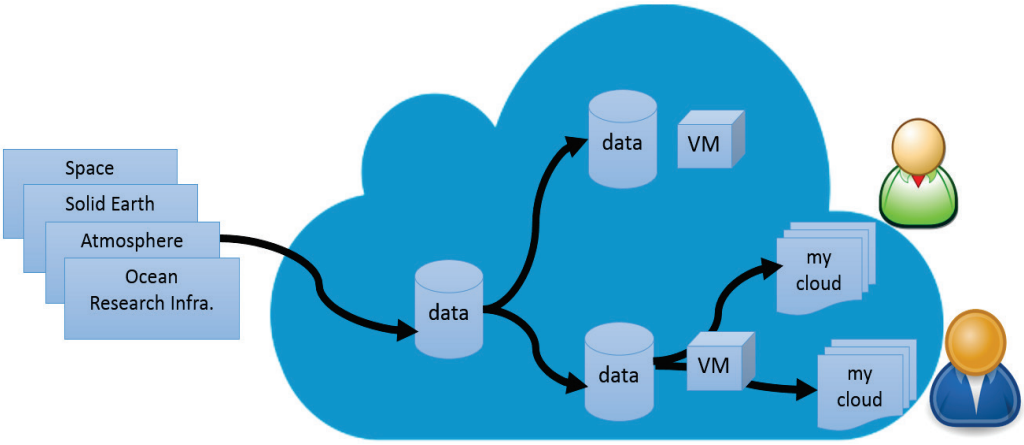

FIGURE 5. BROADCASTING DATA FROM RIS TO THE E-INFRASTRUCTURE VIRTUAL RESEARCH ENVIRONMENT.

Plans for engaging the stakeholders

Workshops and other meetings

In order to engage the stakeholders, workshops has been led and demos will be held, e.g. in the framework of the regularly organised ENVRIweek collaboration meetings, or relevant workshops organized by other communities.

The initial prototype of the use case has been successfully demonstrated during the mid-term review of the project[10]. The experimental results will also be presented in the 8th International Workshop on Data-Intensive Computing in the Clouds in the context of IEEE Supercomputing conference. We will also demonstrate the data flow to push data on a common repository then the integration of Euro-Argo User Interface with the Data Subscription Service.

Once solution has been evaluated, other close stakeholders will be involved (SeaDataNet, SOCAT and EMSO).

Further demonstrations will be focused on interoperability with other communities, arguing on the benefits for all stakeholders involved. Scientific end-users get a subscriptions service to ensure their analyses and models meets the late data available. Generic initiatives will be interested to have a trusted place to synchronise their data collections, providing new features, more visibility. We will have to convince individual researchers of those advantages and the opportunity to use large scale cluster computation. Publishing comprehensive and user-friendly documentation will be a key element.

Questionnaires on requirements

The objective is to provide a solution to other research infrastructure that needs few efforts of integration. Data storage is furnished by EUDAT and computation resources by EGI. An interested infrastructure will have to push their data collection on EUDAT and integrate the Subscription HTTP API.

Criteria for evaluation of usability and degree of "operationality"

The solution is in progress of integration on a data discovery portal and could serve as a showcase.

[10] See https://www.youtube.com/watch?v=PKU_JcmSskw for a recorded demonstration.

APPENDIX D: USING VRES TO SUPPORT USER-DRIVEN DATA ANALYSIS (MECHANISM 3)

Contributed by Ingemar Häggström, Carl-Fredrik Enell, Leonardo Candela, Yin Chen, Baptiste Grenier and Domenico Vitale

Summary

Motivation

For the ENVRIplus RIs, the ultimate goal is to provide quality-checked and calibrated observational data to their user communities. The extent to which an RI supports its users to access data easily and to deliver good and citable research is directly linked with the impact and value of the RI, and thus this issue is given much emphasis when RIs design their ICT systems. In recent years, Virtual Research Environments (VREs) have emerged as an important approach to providing web-based systems to help researchers collaborate. WP7 (T7.1) has set up a VRE for the ENVRIplus community (link to ENVRIplus VRE) using the D4Science platform.

Three WP9 use cases (IC_3, IC_13 and SC_3) are evaluating this service. Detailed descriptions of these use cases have been provided in D9.1 [Chen 2017]. In summary, IC_3 aims to support individual scientists from the EISCAT 3D community to process radar data using their own algorithms. In a similar way, SC_3 aims to enable LifeWatch researchers to develop and share data analysis workflows. Likewise, the objective of IC_13 is to optimise the processing of the eddy covariance (EC) data in order to establish a service that can be used by different RIs that use this micrometeorological technique to measure exchanges of greenhouse gases and energy between terrestrial ecosystems and atmosphere.

Technologies involved:

D4Science https://www.d4science.org/ supports a flexible and agile application development model based on the notion of Platform as a Service (PaaS), in which components may be bound instantly at the time they are needed. In this way, it enables user communities to define their

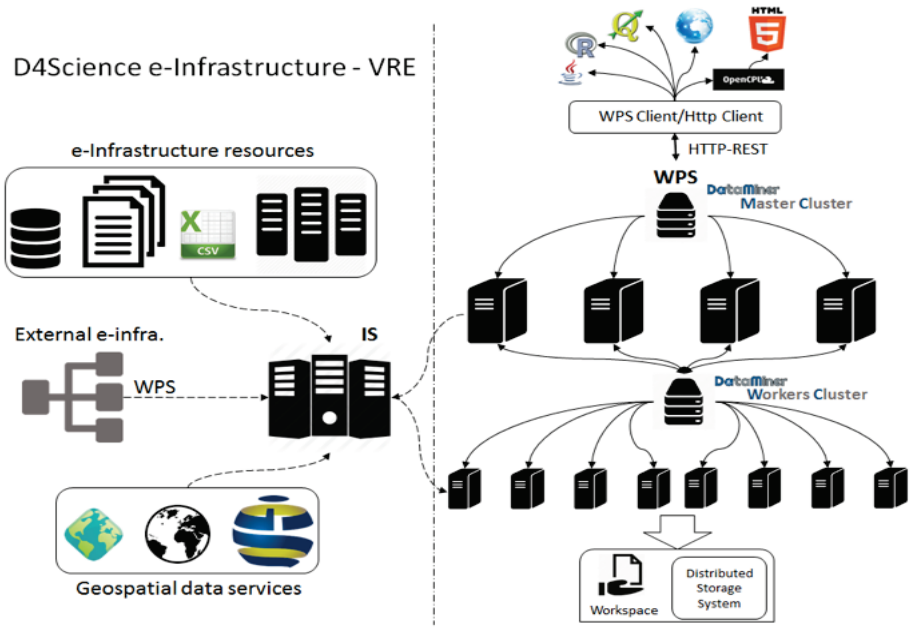

FIGURE 6. THE D4SCIENCE VRE ARCHITECTURE. FROM [CANDELA 2017].

own research environments by selecting the constituents (the services, the data collections, the machines) among the pools of resources made available through the D4Science e-Infrastructure. The architecture is shown in **Figure 6**.

Its front end (left part of **Figure 6**) is the VRE platform that consists of a data analytics framework and a shared workspace where the research objected from the analytics tasks are automatically stored with rich metadata. Object in the workspace can be shared with co-workers. Its back end (right part of **Figure 6**) is the DataMiner service that operates the e-Infrastructure clusters.

As shown in **Figure 7** below, a web GUI is provided that provides a research environment for scientist to find, share, generate, modify and validate data analysis algorithms or workflows.

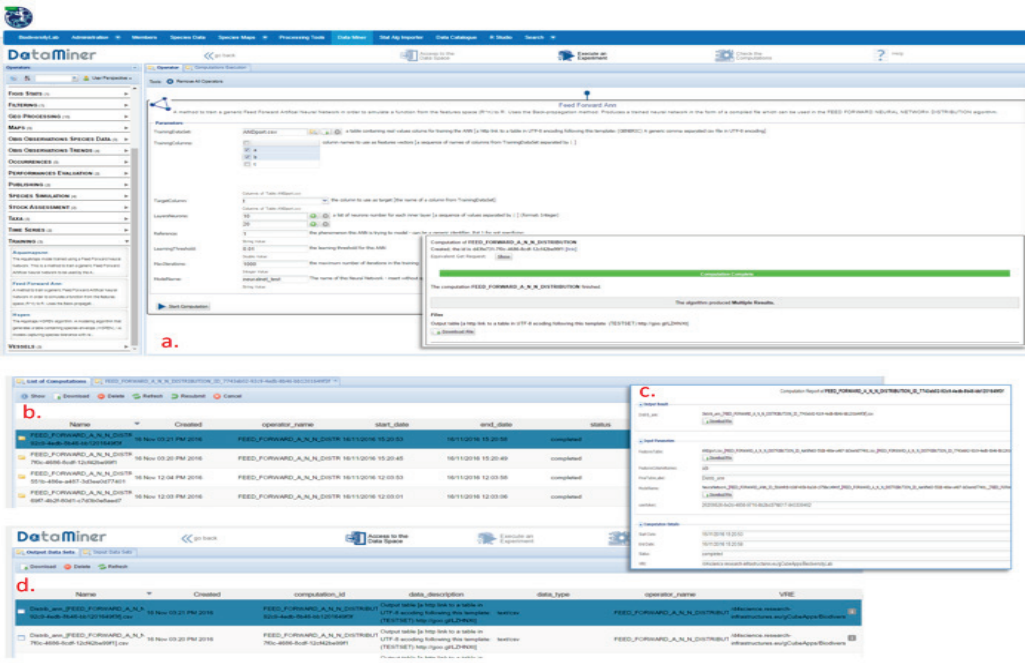

FIGURE 7. THE GRAPHICAL USER INTERFACE (GUI) OF THE D4SCIENCE VRE INTERFACE. FROM [CANDELA 2017].

Goals related to improving

At the technology level, the three ENVRIplus use cases present different integration scenarios.

- IC_3 needs to translate domain specific analysis algorithms originally written in Matlab into Octave, an open-source software. This requires the VRE service to be open and extensible to integrate the new software.
- IC_13 needs to embed the VRE service into a workflow that invokes data and process from different distributing sites within a single RI. This requires the VRE service to be flexible, e.g., following modular design principles, and providing various programmatic interfaces.
- On the other hand, SC_3 needs to integrate the VRE service into a process across different RIs or e-Infrastructures. This gives more challenges in interoperable access that requests the VRE service to have standardised interfaces supporting interoperability.

These usage scenarios are common to ENVRIplus RIs, and the use case Experiences can be shared with them.

Data & metadata access for human users

- EISCAT_3D: Redundant archives with sufficient sustained transfer speeds to handle archival of several petabytes per year. Metadata harvestable in standard formats including OIL-E and/or ISO standards recommended by ENVRIplus, preferably not too many different standards to maintain. Currently (in the legacy EISCAT systems) the catalogues are in SQL databases in our own format.
- IC_13: Provide a user-friendly tool allowing the download of available EC data and metadata in a selected temporal window and in a specified geographical area.

Data & metadata access for computer-driven processing

- EISCAT: Redundant archives with sufficient sustained transfer speeds to handle archival of several petabytes per year. Metadata harvestable in standard formats, preferably not too many different standards to maintain. Currently there are own catalogues and a couple of XML standards, including the OIL-E ontology, which is based on the ENVRI Reference Model.
- IC_13: Processing EC data requires the availability of (i) high-frequency raw data sampled at 10 or 20 Hz and collected in half-hourly files, and of (ii) metadata (both about the data format and instrumental EC tower setup). For ICOS-RI this information is stored at different repositories (i.e. ICOS Carbon Portal, ICOS-Ecosystem Thematic Centre). For computer-driven processes it is therefore essential to develop an automated tool able to link data and metadata. For other RIs not compliant with ICOS requirements, it will be required to standardise raw and metadata files in accordance with the requirement of the processing tool. This standardization step will contribute also to facilitate data sharing, re-distribution and re-use in future.

Context:

Couplings to other ENVRIplus activities

Although the implementations of the three use cases are under WP9 efforts, cross-collaboration has been established with other WPs including, WP5, WP7, and WP8. The key service they test is D4Science which is developed by T7.1. Catalogues services/principles developed by WP8 are also integrated, e.g., by use case IC_3. IC_3 also evaluates Reference Model concepts on metadata aspects.

Similar projects outside of ENVRIplus?

EISCAT data processing similar to the IC_3 case has been considered in both the EGI-Engage Competence Centre for EISCAT_3D (EISCAT_3D-CC) data portal project and an EUDAT pilot project.

The EISCAT_3D-CC portal (https://dirac.egi.eu:9443/DIRAC/) uses DIRAC interware to retrieve existing data from the EISCAT radars and submit jobs for processing on EGI resources. Displaying raw data has been successfully demonstrated.

In the EUDAT pilot project, the data will be made searchable through B2FIND (http://b2find.eudat.eu/) and for data distribution we use B2SHARE (https://b2share.eudat.eu/) and B2SAFE (iRods based storage at CSC).

These developments are now joining forces in the EOSC-Hub, a H2020 project that will start 2018.

Importance

Outside of the current domain (atmosphere, ecosystems, marine, solid earth)

VRE for user-driven data analysis is an commonly requested area not only to environmental scientific community. The implementations from the three use cases can be easily extended to support other domain applications' needs. For example, feedback from agile team IC_3, their implementation would benefit space and solar-terrestrial physics, solar system physics (meteors and asteroids), and astronomy.

In the case of IC_13, the availability of standardised and quality checked EC data product is essential for many purposes: (i) for accurately quantifying the carbon budget, (ii) for better understanding the complex biogeochemical and ecological processes, and (iii) for the verification and tuning of global climate models, mesoscale and weather models, and remote sensing estimates from satellites and aircraft.

Regionally (Europe)

All three use cases come from ESFRI RIs, ICOS, EISCAT_3D and LifeWatch, that have collaborations across Europe. The implementation will benefit their users in the member countries. For example, the investigation in the use case, IC_3, will mainly benefit the European EISCAT member countries, Finland, France, Norway, Sweden, USA and UK, but also data discovery from other countries.

Globally

The collaborations of the involved RIs mentioned in this aspect, indeed beyond Europe level. For example, in the case of IC_3, the global wise collaborations reach China, Japan, Korea, etc. In the case of IC_13, the availability of a computationally efficient VRE designed to the processing of EC data is of interest for ICOS and any other users or RIs using EC data (e.g. AnaEE, eLTER). A standardised processing guarantees homogeneous data more suitable in comparison studies involving different geographical areas.

Target group

Environmental/climate/Earth science domains

There are connections to the following ENVRIplus service implementation cases:
IC_3:

- Science councils
- NRENs and HPC centres
- Academic institutes: space and atmospheric sciences
- Large initiatives: as an example the European Space Agency space weather awareness and space debris programmes will benefit from the services considered in this use case

IC_13:

- Any users and RIs collecting EC data and interested in processing high frequency time series (10 or 20 Hz) through a computationally efficient tool.
- Any users, RIs and networks (e.g. weather forecasting agencies, and global network such as FLUXNET) interested in both long-term and NRT EC data product (i.e. quality checked half-hourly time series).

Plans for engaging the stakeholders

Workshops and other meetings

ENVRIweek meetings provide an important platform for community communication and networking. In the pass ENVRI events, use cases sessions have been organised to present agile implementation results. We will continue to make use of the opportunities to engage the stakeholders. Demonstrations of VRE usability will be presented in the upcoming ENVRIweek. For example, IC_13 agile team is preparing demos that explain detailed steps: 1) EC data and metadata uploading; (2) high-frequency (10 or 20 Hz) EC data processing; (3) quality control and uncertainty quantification; (4) NRT data product. The focus will be on the interoperability with ICOS-RI, with generic users and with other RIs. The demos will show the computational advantages derived from the use of the VRE, and will be accomplished with a comprehensive and user-friendly documentation.

Agile teams are efficient channels to reach out to RIs. Presentations and talks will be given during community conferences and events. For example, members of the agile team for the IC_3 use case have presented e.g. data processing issues during annual EISCAT_3D user meetings and biannual EISCAT symposia. They will also discuss the technology with related partners and projects.

Questionnaires on requirements

Agile teams will help T9.2 to reach RIs, e.g., distribute questionnaires on requirements and collect feedback.

Criteria for evaluation of usability and degree of “operationality”

We will track the usability of the pilot results and evaluate the degree of “operationality”. For example, we will collection the following information from the stakeholders:

- potential user numbers
- measure of Technology Readiness Level (TRL)
- resources (computing, storage, staffing, funding) availability for continuing operation
- agreements between service/data/computing/storage providers

APPENDIX E. ENSURING INTEROPERABILITY OF DATA AND METADATA (MECHANISM 4)

Contributed by Margareta Hellström and Alex Vermeulen, with additional input from Keith Jeffrey and Paul Martin

Summary

Motivation

Improving data access is not only about setting up efficient channels for exchanging digital data objects, for example by implementing user-friendly search interfaces and download services. Indeed, without ensuring that all relevant metadata - administrative, provenance-related, descriptive etc. - is just as easily accessible and interpretable by the end user, the data can be considered more or less useless.

The FAIR guiding principles for data, developed by FORCE11, describe how to make published data Findable, Accessible, Interoperable, and Reusable for potential users [FORCE11 2014, Wilkinson 2016]. Findability means making the data possible to find by potential users, e.g. by describing the data with rich metadata. Accessibility means that the data and metadata should be usable in formats that are understandable by humans and machines, e.g. by adding machine-actionable PIDs. Interoperability pertains to using a metadata scheme that is open and well-defined. Reusability means that the metadata are verifiable, machine-readable and can be used to make proper citations [FORCE11 2014].

By following the FAIR principles, the "7-R's" of data can be fulfilled, i.e. that data are reusable, repurposable, repeatable, reproducible, replayable, referenceable, and respectful [Bechhofer 2013].

Technologies involved

Interoperability involves several components of data management (as outlined by [Atkinson 2016]): Cataloguing, Semantic information linking, Data identification & citation and Processing.

Cataloguing: ideally using international standards for both back end and front end, i.e. based on well-tested metadata storage systems, content organised according to international standards, and accessible via interfaces (web pages, APIs) supporting standardised protocols. "Technologies" for cataloguing include CKAN, CERIF and RDF triple stores.

Semantic information linking: To support both efficient translation between different metadata standards at the catalogue/portal level, and unambiguous interpretation of metadata by the end user, semantic technologies are required. Open information linking for Environmental RIs (OIL-e) [Martin 2015] is a framework developed in theme2 based on the ontological representation of ENVRI RM and mapping (to be mapped) between RM and other metadata standards, including CERIF, PROV and ISO19134.

Data identification & citation: both data and metadata objects should be tagged with unique persistent identifiers, to enable unambiguous references (and citation), resolution and retrieval. The most commonly used PID type for environmental research data is the Handle System, but other PID technologies are also possible.

Processing: it must be possible for (automated) machine-based workflow to easily access data objects, as well as to implement algorithms that interpret related metadata to inform the analysis process. Commonly used workflow engines include Taverna and Kepler. One can also include advanced VRE platforms under this heading.

Goals related to improving

Data & metadata access for human users and computer-driven processing

As discussed in [Jeffrey 2017], a number of conditions must be fulfilled in order to support data use. Specifically, there is a need for commonality of metadata elements across curation, provenance, and cataloguing. This implies that a common core metadata scheme should be used for interoperability – possibly with extensions for particular domains where interoperability is not required. Such a common standard for ENVRIplus has been formulated by WP 8.

Several challenges remain, however, including:

- Metadata collection is expensive so incremental collection along the workflow is required: workflow systems should be evolved to accomplish this and scientific methods and data management working practices should be formalised using such workflows to reduce chores and risks of error as well as to gather the metadata required for curation;
- Automated metadata extraction from digital objects shows promise but production system readiness is some years away. However, metadata provision from equipment-generated streamed data is available already now.

Context:

Couplings to other ENVRIplus activities

The interoperability of research data is an integral part of the ENVRIplus Theme 2 activities: semantic information linking is covered by WP 5, data identification and persistent identifiers by WP 6, processing and workflows by WP 7 and cataloguing by WP 8.

Relevant WP 9 use cases include

Similar projects outside of ENVRIplus?

There are many, including various Research Data Alliance (RDA) working and interest groups, as well as initiatives driven by organisations such as DataONE, ICSU-WDS and CODATA.

Importance

As stated explicitly in the description of Theme 2 [ENVRIplus 2014], environmental research infrastructures are expected to become important pillars not only for supporting their own communities, but also (a) for inter-disciplinary research and (b) for the European Earth Observation Program Copernicus, which in its turn is a main contributor to the Global Earth Observation System of Systems (GEOSS). To ensure that the ENVRIs can fulfil these roles, it is very important that all data-related activities within ENVRIPLUS are well integrated. This requires not only common policies, models and e-infrastructure, but very importantly also concerted efforts to 1) re-harmonise and integrate services and 2) enforce interoperability of data, applications and other services.

Outside of the current domain (atmosphere, ecosystems, marine, solid earth)

Interoperability is important for all scientific domains! In fact, it may be argued that ensuring that data sets are properly described, following standardised vocabularies and information schemata, becomes even more important when data are being (re-)used across domain boundaries, e.g. when a health researcher wants to incorporate data on air pollution as background material for her own clinical studies. If the air quality data isn't well documented, serious mistakes and errors could ensue if the collected information is used in the wrong way by the medical professional. In addition, if the air data is provided only in some proprietary format, the dataset may be inaccessible unless the same software package is made available to the health researcher.

Regionally (Europe) and globally

Funders and regulators in different regions may choose to impose specific rules or recommendations concerning e.g. specific metadata standards that should be adhered to. One example is the INSPIRE Directive from the European Union, which "aims to create a European Union spatial data infrastructure for the purposes of EU environmental policies and policies or activities which may have an impact on the environment" [INSPIRE 2017].

Target group

Stakeholders

Authorities, governmental agencies etc.

Funders are pushing strongly for research data (especially publically funded) to be made reusable -- both by other practitioners in the original science domain and also, when possible, by researchers from other fields. This provides a strong argument for making sure that data are by default made FAIR as part of the normal data management routine. This is indeed a reasonable demand, as many important research discoveries have been made by re-working old data and/or by comparison of old data with recently collected data. This is particularly true in the Earth sciences, where understanding processes taking place in any of the "spheres" (atmosphere, biosphere, cryosphere, hydrosphere and geosphere) usually requires long-term observation and subsequent analysis of data from multiple sources. In addition, validation and re-validation of research results also require open and understandable access to the data used in the preparation of the original publication [Jeffrey 2017].

Large “generic” initiatives

Many of the large "generic" initiatives like Copernicus and GEOSS are operating and/or planning for monitoring and prediction services on a regional or global scale. These services require access to data on e.g. concentrations and fluxes of greenhouse gases, pollutant and aerosol loadings, or other environmental variables from many different sources – some of which are ENVRIplus partners. Successful operation obviously demands that data are easily accessible in a timely manner, that the transfer is expedient and that all data both are traceable, persistently identified, and follow strictly regulated standards with respect to content type.

User communities

End users of ENVRIplus data products include both other RIs and environmental and Earth science researchers from across Europe and the rest of the world. Ensuring that all data products,

ranging from near real-time data to fully quality controlled finalised datasets, are FAIR is a prerequisite for earning users' trust and instilling their confidence in ENVRIplus.

Environmental/climate/Earth science domains

See the previous point.

Plans for engaging the stakeholders

Questionnaires on requirements

As part of WP 5 activities, questionnaires were sent out in 2015-2016 to all ENVRIplus partners to map out their requirements on the interoperability-related topics of Cataloguing, Semantic information linking, Data identification & citation and Processing. The outcome of this study was reported in deliverable D5.1 [Atkinson 2016].

Criteria for evaluation of usability and degree of "operationality"

The FAIR principles should form the basis for defining suitable evaluation criteria related to interoperability. A dedicated FAIRness evaluation and grading system is under development [Doorn 2016], and could be applied also by WP9.

Workshops and other meetings

In the framework of the November 2016 ENVRIplus Week (Prague, Czech Republic), a workshop was co-organised with Copernicus. The rationale for the meeting was that although many environmental RIs are placing Copernicus on the top-list of their users, very few ENVRIs have clearly formalised specific agreements with Copernicus, directly or indirectly, for any service provision. The workshop provided an opportunity for several RIs across the different domains and at different level of maturity to expose the current status of their interaction with Copernicus. Equally important, representatives from Copernicus services and the EEA were able to outline their expectations on timeliness of data transfer, data quality, and interoperability aspects.

The possibilities to arrange further workshops with Copernicus and other generic initiatives during the remaining ENVRIplus weeks are under investigation. Other options for stakeholder interactions include giving presentations about ENVRIplus data products at e.g. Copernicus collaboration meetings as well as at international and national scientific conferences.