Ref. Ares(2017)4273936 - 01/09/2017

D9.1
Service deployment in computing and internal e-Infrastructures

WORK PACKAGE 9 – Service Validation and Deployment

LEADING BENEFICIARY: EGI FOUNDATION

Author(s):	**Beneficiary/Institution**
Yin Chen & Baptiste Grenier	EGI Foundation
Margareta Hellström & Alex Vermeulen	Lund University
Markus Stocker & Robert Huber	University of Bremen
Barbara Magagna	Umweltbundesamt Austria
Ingemar Häggström	EISCAT
Markus Fiebig	NILU
Paul Martin	UvA
Domenico Vitale	Uni. Tuscia
Glenn Judeau & Thierry Carval	IFREMER/EuroArgo
Thomas Loubrieu	IFREMER/SeaDataNet
Abraham Nieva	Cardiff University
Keith Jeffery	NERC
Leonardo Candela	ISTI-CNR
Jani Heikkinen	CSC

Accepted by: Yannick Légre (WP 9 leader), Zhiming Zhao (Theme2 Leader)

Deliverable type: [REPORT]

Dissemination level: PUBLIC

Deliverable due date: 30.8.2017/M28

Actual Date of Submission: 31.8.2017/M28

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

ABSTRACT

This deliverable reports the group efforts of Working Package 9 task T9.1 on service integration and deployment during May 2016 (M13) to Aug 2017 (M28). It introduces the process and method of identifying community use cases for testing and validating ENVRIplus service solutions. It provides the information details of the agile investigations. Furthermore, it discusses the plans and status of WP6-8 for service deployment and introduces deployment approaches using e-Infrastructures. Finally, the success metrics for evaluating the implementation results of the use cases are also introduced.

Project internal reviewer(s):

Project internal reviewer(s):	**Beneficiary/Institution**
Ari Asmi	University of Helsinki
Malcolm Atkinson	The University of Edinburgh
Wouter Los	University of Amsterdam

Document history:

Date	**Version**
1.6.2017	Draft for comments
30.7.2017	Completed version for internal review
31.8.2017	Accepted by Theme2 leaders

DOCUMENT AMENDMENT PROCEDURE

Amendments, comments and suggestions should be sent to the authors (Yin Chen yin.chen@egi.eu)

TERMINOLOGY

A complete project glossary is provided online here:

https://envriplus.manageprojects.com/s/text-documents/LFCMXHHCwS5hh

PROJECT SUMMARY

ENVRIplus is a Horizon 2020 project bringing together Environmental and Earth System Research Infrastructures, projects and networks 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 harmonize 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 structuring 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.

TABLE OF CONTENTS

EXECUTIVE SUMMARY

ENVRIplus services developed in Theme2 WP6-8 are valid only if positively evaluated by ENVRIplus participating Research Infrastructures (RIs). As the major task of WP9, T9.1 undertakes testing and validation of ENVRIplus service solutions in realistic usage scenarios.

This document builds on the progress in T9.1 during the period May 2016 – August 2017. At the outset, a method of jointly collecting and identifying appropriate community use cases was introduced. 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.

An agile approach is used for the implementation of the selected use cases. Implementation teams follow agile principles for service integration and testing. Most of the team members are researchers directly involved in ENVRIplus RIs, with good knowledge of respective community requirements, and through their agile activities these teams help establish interdisciplinary collaboration across the ENVRIplus community.

The agile implementation teams are self-motivated and work efficiently, and several have already made relevant progress. A showcase session was organized by WP9 during the 4th ENVRIweek, Grenoble, May 2017, where 4 use cases gave demonstrations to the entire ENVRIplus RIs community. One use case was selected as one of three Theme2 highlights and presented in the project review meetings.

Task T9.1 supports the service deployment for the ENVRIplus services developed in Theme2 WP6-8, and for the integration of services developed through the use cases. The respective deployment plans and requirements were discussed with WP6-8 and with the use case implementation teams. Where needed, e-Infrastructure resources were provided for service testing. As several use cases utilize the EGI Federated Cloud[1], a guideline was developed explaining the deployment approaches on this e-Infrastructure. This guideline is now available[2] to assist ENVRIplus service providers in understanding and utilizing these technologies.

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 is introduced together with a scoring method. This T9.1 activity is closely linked to T9.2 that facilitates and encourages the adoption of ENVRIplus Theme2 results.

In future, T9.1 will continue to coordinate the use case developments towards demonstrators. We will work closely with T9.2 to promote the ENVRIplus service solutions, and in particular explore how to connect ENVRIplus services with the European Open Science Cloud (EOSC)[3]. Two European leading e-Infrastructures (EGI and EUDAT) are involved in WP9, and they provide important support to help ENVRIplus RIs exploit the usage of e-Infrastructure resources and services. These activities and experiences will facilitate the positioning of ENVRIplus RIs within the EOSC.

[1] EGI Federated Cloud: https://www.egi.eu/federation/egi-federated-cloud/
[2] Guideline for Service Deployment: https://confluence.egi.eu/display/EC/Guideline+for+Service+Deployment
[3] European Open Science Cloud (EOSC): https://ec.europa.eu/research/openscience/index.cfm?pg=open-science-cloud

1 INTRODUCTION

The objective of Working Package 9 (WP9), as defined by the DoA, is *"to validate the results of ENVRIplus Theme2 development and deploy the developed software onto computing and data infrastructure and investigate how to operate developed services within RIs"*. This is an important step to ensure services developed by ENVRIplus Theme2 are relevant to the user community and are useful to resolve real-world problems.

WP9 defines two tasks. Among them, Task 9.1 (T9.1) *"focuses on the technical issues of software validation, integration and release management, and of deploying developed results on computing and data infrastructure"*. To achieve these, according to the DoA, *"a complete use case will be defined and its full implementation analysed and validated. The use case will be selected to preferably involve several research infrastructures, have the interest and participation of active scientific communities, and have a clear impact."* This is an approach to test Theme2 services in realistic usage scenarios and evaluate their functional and non-functional features against various user requirements.

Since its start in May 2016 (M13), T9.1 has been endeavouring on collecting community use cases and organizing agile activities to implement those use cases. The primary challenges have been how to maximize the involvement of ENVRIplus partner Research Infrastructures, how to ensure use cases of high quality, and how to efficiently implement those use cases. We have set up a process to call for use case submission from the entire ENVRIplus community, and invited experts to review the submitted use cases. The driver is to define appropriate cases that address community's real needs, implement ENVRIplus results in the relevant RIs, and demonstrate their value.

Use cases are implemented following an agile approach. 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 [Atkinson 2015]. For each use case, an agile team is composed. The agile teams are self-motivated, flexible and efficient. They respond to test various technologies, integrate ENVRIplus services with community services to fulfil the investigation goals, and to enable the final usage of developed services by testing and deploying them onto e-Infrastructures.

E-Infrastructure is a term that designates the new generation of Information and Communication Technology (ICT)-based infrastructures. They typically exploit several components and layers, such as networks, supercomputers and other computing resources, storage, software, remote resources and instrumentation, e.g., sensors. Such elements are seamlessly interconnected and can be access by users all around the world, regardless of their geographical location. E-Infrastructures allow scientists to share information securely, analyse data efficiently, and collaborate with colleagues worldwide. They are an essential part of modern scientific research and a driver for economic growth. The European Commission is heavily investing in e-Infrastructures through its H2020 Programme, and these new platforms are by now considered to be key enablers of the European Research Area (ERA).

Two Europe leading e-Infrastructures join in WP9, namely EGI and EUDAT. EGI provides European and global federation of computing and storage resources. EUDAT offers an interoperable layer of common data management services for researchers and research communities. Technical experts from these e-Infrastructures participate in the agile activities and provide important support to help ENVRIplus RIs exploit the usage of e-Infrastructure resources and services technology.

Many ENVRIplus community members are not familiar with e-Infrastructure technology. In order to help Theme2 service development teams understand the methods and approach to deploy services onto e-Infrastructures, T9.1 provides technical support through presentations and demonstrations. A guideline[1] is also provided that describes how to deploy services on a concrete e-Infrastructure computing environment, namely EGI Federated Cloud[1]. Section 4 reports this activity. It also discusses the plans and status of WP6-8 for service deployments using e-Infrastructures.

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. If the requirements or the expectations are not satisfied, it is important to understand the reasons. An initial evaluation metric is introduced (in Section 5), together with a method defining how to assign scores. 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.

The rest of the report is organized as follows:

- Section 2 describes the approach and process for use case collection;
- Section 3 provides detailed information about use cases and agile activities in service integration and deployment;
- Section 4 reports the activities on service deployment;
- Section 5 describes the approach to service evaluation;
- Section 6 summarizes the report and discusses future actions.

2 METHODOLOGY

2.1 Rationale and Terminology

Since ENVRIplus Work Packages (WPs) and deliverables were defined collaboratively by Research Infrastructures (RIs), WPs results should be relevant to most RIs. To assure that the RIs adopt results, it is important that they indicate their interest already at an early stage of development; ideally they should become engaged and steer work to meet their needs through interdisciplinary co-design and co-development. WP9 follows this approach to analyze the RIs interests and requirements, and to identify use cases for understanding user demands, defining the requirements, and develop demonstrators.

Since the ENVRIplus community involves people from different backgrounds, in particular scientific researchers, ICT developers, etc. The term *Use Case* means different things to different people. To clarify the terminology in this context, we describe the conditions and implications for each of the three different cases: *science*, *test*, and *implementation*.

2.1.1 Science case (SC)

Conditions

- A research project dependent on various new RI services (methods)
- Cutting edge international research with excellent scientists
- The science case serves as a selling story for the ENVRIplus project

Implications

- The use case leader should be willing to serve as a pilot and test solutions
- The use case is developed in one of the ENVRIplus partner research organizations
- Funding to be provided for a researcher/postdoc in the science case

2.1.2 Test case (TC)

Conditions

- The test case is built on a new and developing RI service, so that ENVRIplus staff can understand its rationale
- The test case covers topics of relevance to various WPs, such as instrumentation, data flows, and training
- The test case is part of the RI's portfolio of implementation cases

Implications

- The involved RIs have enough developers to work together with ENVRIplus staff
- The test developments are to be shared with ENVRIplus partners

2.1.3 Implementation case (IC)

Conditions

- Each RI describes its portfolio of new and/or enhanced services that they expect from ENVRIplus results. The descriptions can initially be general, derived from the ENVRIplus WPs
- ENVRIplus staff work with the RIs on these descriptions, which in the course of the project will be gradually updated with more details

Implications

- Implementation cases selected and adopted by interested RIs
- Both RIs and ENVRIplus invest in the actual implementation and associated services

2.1.4 Identification of cases

The driver in the process is to define appropriate cases that implement ENVRIplus results in the relevant RIs. A plan should start with describing the desired *implementation cases* in these RIs and then select some *test cases* that will assist in discovering best practices for full implementation. Ultimately these should be embedded in one or more *science cases* with scientific users.

Identification of Implementation cases

Theme2 (WP5) has worked with the RIs in describing their portfolio of required new and/or enhanced services, resulting in the implementation cases. This is essentially the "translation" of the WP/Task description into the needs of each RI. The WP5 requirement collection phase addressed explicitly each of the Theme2 topics with questionnaires based on examples for relevant use case scenarios (within RI and between RIs). This has resulted in implementation plans for each topic and for each interested RI.

Identification of Test cases

The most interesting test cases are the ones that are related to more than a single WP or Theme and may cover some implementation cases. Test cases should be small and time limited, concrete but not necessarily have immediate success stories (science case) from domain scientists. Test cases should help with a) the concrete definition of some testing activities associated to a few RI services, b) the identification of the current gaps in technologies, and c) plans for delivery of results and "lessons learned" in the short term.

Identification of Science cases

Ideally, test cases are embedded in one or more scientific use cases, research projects in need of new developments and services that ENVRIplus may provide. The RI domain leaders may suggest cross-cutting use cases serving as selling stories. However, to get the cooperation of a research project and commitment to make progress, it is required to fund a researcher/postdoc willing to work on the pilot. When such funding is not possible, a more modest approach is to integrate selected test cases and find potentially interested scientists.

2.2 Process for Use Case Collection and Implementation

Figure 1 describes the process for use case collection and implementation. It includes three main steps: 1) Use Case Collection, 2) Use Case Evaluation, and 3) Agile Activities.

2.2.1 Step 1: Use Case Proposals

During the first ENVRIweek, a brainstorming session was organized to involve different RIs, Themes and WPs to discuss possible use cases. In total, 6 Science Cases, 16 Test Cases and 12 Implementation Cases were identified. For each use case, a key contact person was appointed for the initial analysis and assessment of implementation feasibility.

After the meeting, a formal call for the submission of use cases was announced. The proposers were asked to complete a Case Description Template, designed to include ENVRI Reference Model[4] terms. The template includes the following aspects (see Appendix A):

- Objectives and scope
- Expected results and performance
- Involved RIs
- Relevance to WPs and Tasks

---

[4] ENVRI Reference Model: https://confluence.egi.eu/display/EC/ENVRI+Reference+Model

- Activity plan and time table
- Allocation of resources (staff time and budgets) from RIs and ENVRIplus
- Feasibility assessment

By the first deadline (after one month), 1 Science Case, 3 Test Cases and 8 Implementation Cases were submitted. The call was kept open for another 3 months to allow for more submissions. We stopped collecting new use cases when the project entered the design stage.

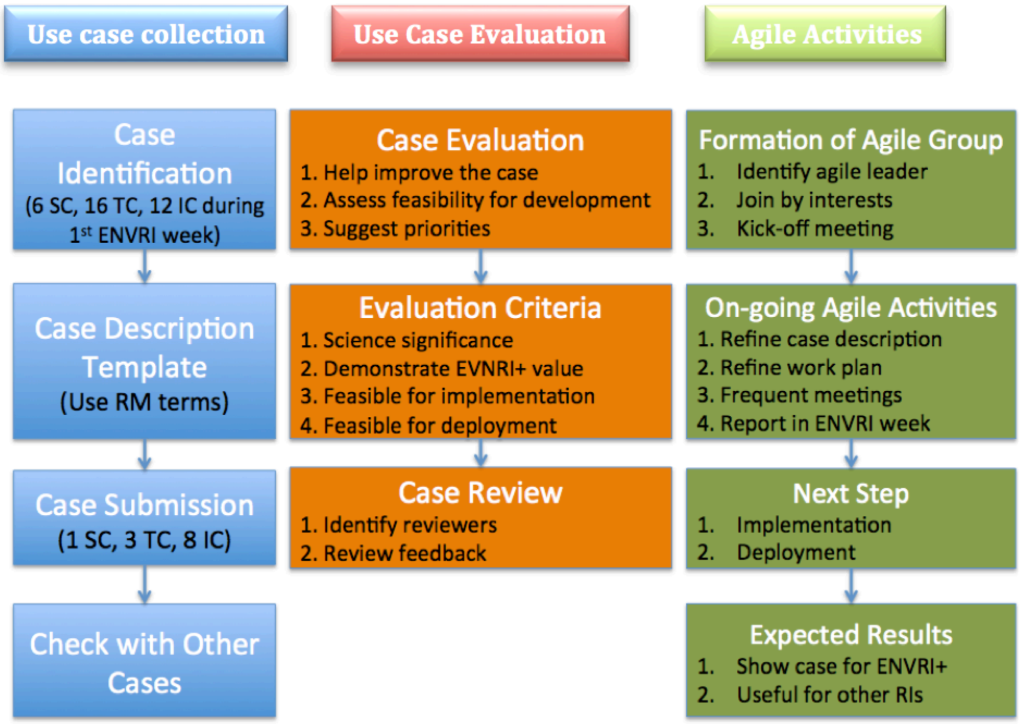

FIGURE 1. USE CASE COLLECTION PROCESS

2.2.2 Step 2: Use Case Proposal review

The purpose of use case review and evaluation is to help proposers improve the description of their use cases, by assessing the scientific significance, alignments with the development agenda of Theme2, and feasibility of development and deployment. After the review, some case refinements may be required, and a set of excellent use cases was recommended for further implementation with higher priority.

An excellent use case needs:

- To be feasible for implementation and deployment
- Well demonstrate the value of ENVRIplus, e.g. support multidisciplinary research and be a selling story for ENVRIplus services
- Have high scientific value
- Reflect strong community interests.

A Use Cases Review Feedback Form was designed that includes the evaluation criteria (see Appendix B).

For each use case, 4 reviewers were identified (on volunteer-base, depending on individual interests) with related background knowledge corresponding to 4 evaluation areas (see Appendix B).

The outputs of the review were reported to the Theme2 development teams for scheduling the implementation activities. Reviewers were also encouraged to advise case proposers and leaders of the development tasks to form agile groups.

2.2.3 Step 3: Set up Agile Activities

An Agile Development method is used for use case development. The Agile Development approach refers to a set of software development methods in which requirements and solutions evolve through collaboration between self-organising, cross-functional teams. It supports promoting adaptive planning, evolutionary development, early delivery, continuous improvement, and encourages rapid and flexible response to change.

There are a number of key activities during this phase:

Formation of the Agile Group

The starting point is to identify the agile group leader who is willingly to lead the group to achieve its goals. The agile group leader's response is to gather interested parties who between them have the required skills, knowledge, experience and connections, and who can allocate sufficient priority time to the task, The leader then needs to establish and maintain effective collaboration between them. There should be a Kick-off meeting to discuss commitment to roles of the agile team's members and to initiate the communication and coordination processes.

On-Going Agile Activities

The agile tasks should start with the refinement of the case descriptions and the work plan. The agile leader should organize frequent virtual meetings to discuss issues and monitor the progress. During ENVRIweeks, a dedicated session is organized for all agile teams to report their progress to the whole ENVRIplus community. Demos should be given. Each agile team is also requested to explain how their results can be made generic and so benefit other ENVRIplus RIs. This will encourage the adoption of ENVRIplus service products. During ENVRIweeks, side meeting rooms and facilities are arranged for agile groups to have face-to-face meetings.

Expected Results

The implementation results of each agile group will be the showcases for ENVRIplus that demonstrates a contribution of the project both internally and externally. When possible, the solutions should be generic so that they may be adopted by other RIs.

2.3 Initial Results

The approach has gained broad community support. Table 1 lists the 13 selected use cases. The numbering in Column 1 retains the numbering of the initial proposal submissions.

TABLE 1. THE 13 USE CASES DISCUSSED IN THE REPORT

#	Use Cases	Agile Group Leaders
SC_3	How do mosquito born diseases emerge and what are the trends?	Matthias Obst (UGOT), Baptiste Grenier (EGI)
TC_2	EuroArgo data subscription service	Thierry Carval (IFREMER), Yin Chen (EGI)
TC_4	Sensor registry	Thomas Bloubrieu (IFREMER)
TC_16	Description of a national marine biodiversity data archive centre	Dan Lear (MBA), Abraham Nieva & Alex Hardisty (CU)
IC_1	Dynamic data identification & citation	Alex Vermeulen & Margareta Hellström (LU)
IC_2	Provenance implementation	Barbara Magagna (EAA)
IC_3	Support EISCAT-3D users to reprocess data using user's algorithms	Ingemar Häggström (EISCAT), Leonardo Candela (ISTI-CNR)
IC_9	Quantitative accounting of Open Data use	Markus Fiebig (NILU)
IC_10	Domain extension of existing thesauri	Barbara Magagna (EAA)

IC_11	Semantic linking framework	Zhiming, Zhao & Paul Martin (UvA) Barbara Magagna (EAA)
IC_12	Implementation of ENVRI RM for EUFAR and LTER	Barbara Magagna (EAA)
IC_13	The eddy covariance fluxes of GHGs	Dario Papale & Domenico Vitale (UNITUS)
IC_14	SOS&SSN ontology based data acquisition and NRT technical innovations	Robert Huber & Markus Stocker (UniHB)

Agile teams are self-motivated, flexible and efficient. More importantly, this approach encourages collaboration across different research infrastructures, work packages and organizations. Table 2 shows the involvement of work packages and ENVRIplus research infrastructures in each use case, illustrating the collaborations established among the ENVRIplus community.

TABLE 2. INVOLVEMENTS OF WORK PACKAGES, ENVRIPLUS RIS AND ORGANIZATIONS

#	WPs	ENVRIplus RIs	Partners
SC_3	WP7, WP9	LifeWatch-SW, EGI, D4Science, BioVel	UGOT, EGI, ISTI-CNR
TC_2	WP7, WP8, WP9	EuroArgo	IFREMER, EGI, EUDAT, UvA
TC_4	WP8, WP5, WP9	EMSO, EuroArgo, EPOS, ICOS, SIOS, ANAEE	IFREMER, CU, Plocan, NERC.BODC, LOCEAN, IPGP, CNR/IBIMER, IPSL, 52°North, MARUM, CSEM, INRN
TC_16	WP5, WP9	DASSH, MEDIN	MBA, CU, UvA
IC_1	WP6, WP9	ICOS, ANAEE, ACTRIS, LTER-Europe, IAGOS, EMSO	ICOS, AnaEE, IAGOS, ACTRIS, UniHB/PANGAEA, EAA
IC_2	WP6, WP8	LTER-Europe, ANAEE, IS-ENES, EPOS, ICOS, ACTRIS	EAA, EUDAT, UvA, INRA, DLRZ, CNRS, ING, LU, CINECA
IC_3	WP7, WP9	EISCAT-3D	EISCAT, ISTI-CNR
IC_9	WP6, WP8	ACTRIS, ICOS, EPOS	NILU, ICOS, ETHZ
IC_10	WP5	LTER-Europe, LifeWatch-Italy, EMBRC	EAA, Università di Salento
IC_11	WP5	All ENVRIplus RIs	UvA, EAA
IC_12	WP5	LTER-Europe, EUFAR	EAA, DLR
IC_13	WP7, WP9	ICOS	ICOS-ETC (University of Tuscia, Viterbo, Italy), ISTI-CNR.
IC_14	WP3, WP9	EMSO, FixO3, EPOS, SIOS, ANAEE, SeaDataNet, EuroArgo	UniHB, Ifremer, CNRS, CNR, EGI, EAA, PANGAEA

2.4 Services in the context of ENVRIplus

2.4.1 Relationships with the Data for Science components

ENVRIplus Theme2 ("Data for Science") is concerned with developing services and models and e-Infrastructure that support environmental RIs across all aspects of the research data lifecycle – from collection of sensor data all the way to finalized data products. The work is being organized according to the model of research data management shown in Figure 2, where research infrastructures are connected to underlying e-Infrastructures on which the envisaged ENVRIplus services and systems will run, via "pillars" representing six distinct Development Research Method (DRM) components: cataloguing, curation, identification & citation, optimization, processing and provenance. The pillars are connected by three “cross beams” representing aspects that are vital to interoperability, such as a system architecture based on a common reference model, and shared services for meta-information exchange.

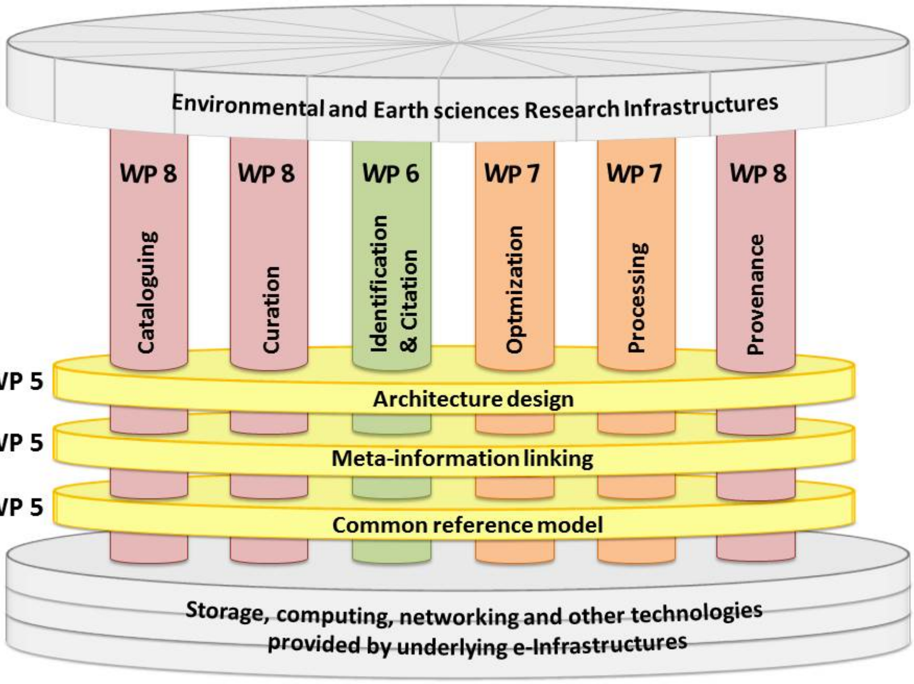

FIGURE 2. THE “PILLARS” AND “CROSS BEAMS” CONNECTING ENVIRONMENTAL AND EARTH SCIENCE RIS WITH E-INFRASTRUCTURE PROVIDERS [ATKINSON 2016].

As a consequence of the often complex functionality required by RIs (as well as most other research institutions and organizations), a majority of the services and data management options under investigation by ENVRIplus are characterized by interdependencies and interconnections between several DRM components. By referencing back to the structure outlined above, it is, however, possible to get a comprehensive overview of the ENVRIplus service development, testing and deployment activities, to help identify commonalities, synergies and gaps between different tasks and cases.

As indicated in Figure 2, the responsibilities for the component topics have been divided among WP6-8, while WP5 leads the work on the cross-cutting topics. Table 3 below shows the involvement of the six “pillars” and the three “cross beams” in each of the science, use and implementation cases introduced in the previous section.

TABLE 3. DISTRIBUTION OF CASES ACROSS THE THEME 2 “PILLARS” AND “CROSS BEAMS”.

Topic	Work Package*	Use Cases
Cataloguing	WP8	TC_2, TC_4, IC_2, IC_8
Curation	WP8	TC_2, TC_4, IC_2
Identification & Citation	WP6	TC_4, IC_1, IC_2, IC_9
Optimization	WP7	TC_2, IC_3, IC_13
Processing	WP7	SC_3, TC_2, IC_3, IC_13, IC_14
Provenance	WP8	TC_2, TC_4, IC_2, IC_8, IC_9
Architecture design	WP5	SC_3
Semantic-information linking	WP5	TC_4, TC_16, IC_10, IC_11, IC_12
Reference model	WP5	TC_16, IC_12

* In addition, WP3 from Theme 1 is involved in IC_14.

2.4.2 Relationship with Theme2 Work Packages

Common to all considered use cases, the agile work obviously benefits strongly from both active involvement by one or several RIs as well as direct connections to concrete situations and implementations. The latter include sensors in use or in development, available data streams or legacy data, factual routines for data processing, existing web services, remote access arrangements, and training programmes. The coordination of the work between the development areas (WP5-8) and the deployment areas (WP9) has to not only stimulate RI engagement, but should preferably be led by those expert in the challenges RIs face.

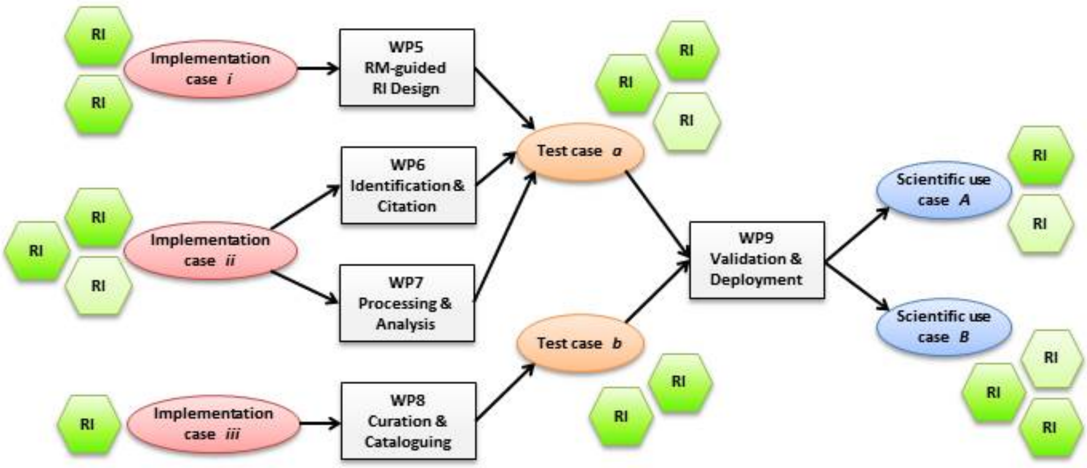

FIGURE 3. THE ROLE OF WORK PACKAGES IN THE OVERALL ORGANIZATION OF CASE MANAGEMENT. NOTE THE COORDINATION ROLE OF WP9, ESPECIALLY ACTIVE AT THE LEVELS OF TEST CASE AND SCIENTIFIC USE CASE MANAGEMENT, WHILE WPS 5-8 ARE MORE INVOLVED IN ORGANIZING AND LEADING THE IMPLEMENTATION CASE DEVELOPMENT.

Figure 3 depicts the general relationships between use/test/implementation cases with the tasks involved in ENVRIplus work packages. Because it is concerned with the development of the cross-bridging DRM components, WP5 has a natural role in several of the implementation and test cases. The ENVRIplus development areas (WP6-8) are, in addition to use cases already defined in their respective DoA, involved in implementation cases that have been identified together with interested RIs. The role of WP9 is to select test cases for deployment, and to coordinate software validation and software integration (how to integrate services with different RIs communities' services and e-Infrastructures' services). Finally, WP9 is also responsible for coordinating the overarching science cases, which build on deployed test case services and other outcomes that are able to support the investigation of realistic research questions that involve both ENVRIplus RIs and their user communities.

3 AGILE ACTIVITIES ON SERVICE INTEGRATION AND VALIDATION

This section provides information about use cases and agile investigations. Emphasis is given to their actual activities on service integration and validation, and achievements up to this stage. The information has been provided directly by the agile teams.

3.1 Use Case SC_3: How do mosquito born diseases emerge and what are the trends?

Group Leaders

Matthias Obst (UGOT), Baptiste Grenier (EGI)

Objectives

The scientific vision of this use case is to enable a more efficient management of mosquito-borne diseases and nuisance mosquitoes. Mosquito-borne infections are among the most important new and emerging diseases globally and in Europe, and in order to predict diseases transmission areas statistical correlation approaches are used.

This objective should be achieved following the technical vision of supporting researchers in combining biological and hydrological data in a collaborative and evolving Virtual Research Environment (VRE) allowing intensive statistical computations: researchers should be able to easily share and use algorithms that they can adapt and use with their own data.

Partners

UGOT, EGI, ISTI-CNR

Related RIs

Swedish LifeWatch, EGI, D4Science, Biodiversity Virtual e-Laboratory (BioVel)

Activities

A demonstration was presented at the 3rd ENVRIweek. The main objective was to validate that d4science.org provided by WP7 (task T7.1) could be used to create a customized VRE for researchers to easily share algorithms, data and allow them to run their computations on e-Infrastructures such as EGI.

This demonstration presented how a researcher can easily upload and integrate an R-based algorithm in d4science, making it available to other researches, members of the VRE in which the algorithm was published.

Once published, researchers can discover the algorithm and use it with their own data. It is also possible to adapt the algorithm and to share improved versions.

When using the algorithm, the computation is done on federated resources, such as those provided by the EGI e-Infrastructures.

Achievements

The main achievement is the demonstration of a proof of concept using d4science.org to setup a community-centric VRE allowing researchers to share a simple algorithm with some data and allowing to run computation on EGI Federated Cloud[1] e-Infrastructure. The data was gathered and produced by the involved RIs and the dynamic computation was executed on EGI Federation's resources. The architecture is depicted in Figure 4.

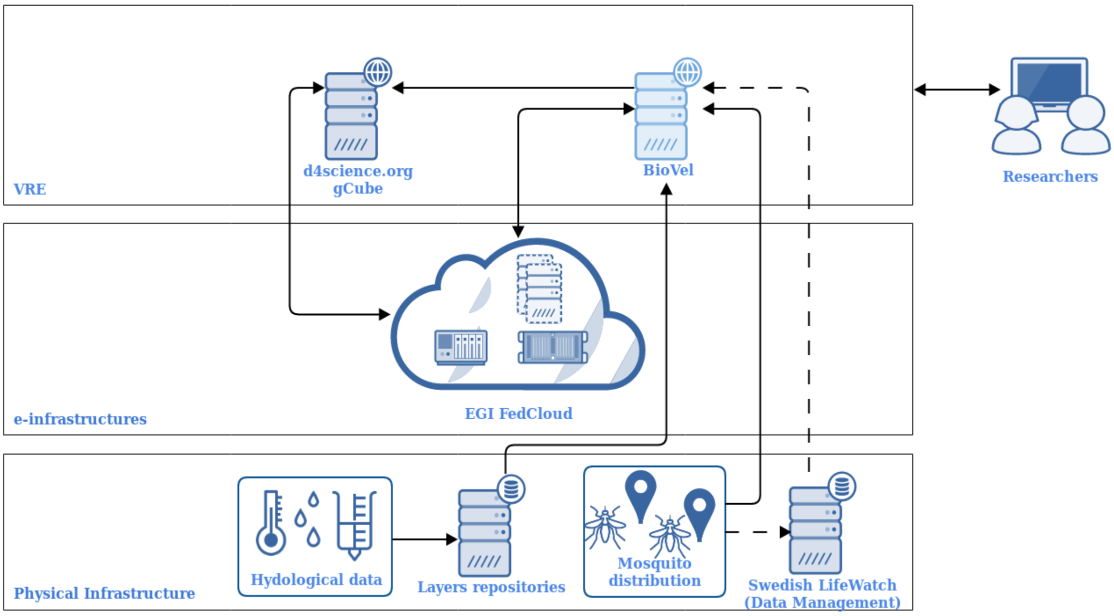

FIGURE 4: MOSQUITO USE CASE ARCHITECTURE

Next Steps

Now that a basic algorithm has been integrated into d4science.org it is necessary to see how more complex algorithms and workflows, potentially integrating remote web services, can be deployed and used by researchers.

It will also be necessary to understand how to adapt and configure the VRE for the specific needs of the researchers.

These two points will allow to more easily see how this use case can be generalized to support different RIs having similar needs.

3.2 Use Case TC_2: EuroArgo data subscription service

Group Leaders

Thierry Carval (IFREMER), Yin Chen (EGI)

Objectives

An extensive number of Research Infrastructures need to publish and give access to datasets that may accumulate over time and need to remain available for download for researchers. The accumulated datasets are queried and analyzed, leading to new data results. Keeping an eye on the accumulated and result datasets is time consuming, thus a subscription model was adopted to facilitate researchers needs. The subscription model does not require direct interactions between the researcher and possibly time consuming analysis actions, thus allowing more flexible design and integration of the system components (Figure 5). In practice this means that even when time consuming actions can be optimized to operate within time-constrained requirements, the adopted model alleviates effects of long-running actions on the data, especially when the size of accumulated datasets increases.

The objective of this test case is to develop and integrate a system to access, download, and subscribe to EuroArgo DataSets. The EuroArgo community aggregates the marine domain datasets into a community repository from which the data is pushed to the EUDAT B2SAFE[5]

[5] EUDAT B2SAFE service: https://www.eudat.eu/b2safe

service. The service allows data registration through data identification. Optionally the community registers their subscription actions and their parameters in B2SAFE[5]. Data could be requested through interfaces provided by EUDAT and IFREMER web site. Subscription is managed by EUDAT and actions are processed using cloud resources, such as EGI Federated Cloud (EGI FedCloud)[1].

Users can select different attributes for their subscription like localization, stations, parameter, update frequency and more.

Partners

EUDAT, EGI, IFREMER, UvA

Related RI

EuroArgo

Activities and Achievements

IFREMER provided accumulating monthly marine domain datasets to be pushed to EUDAT e-Infrastructure (B2SAFE[5]). The datasets were then synchronized between EUDAT and Cloud resources, e.g. EGI Federated Cloud[1], on demand.

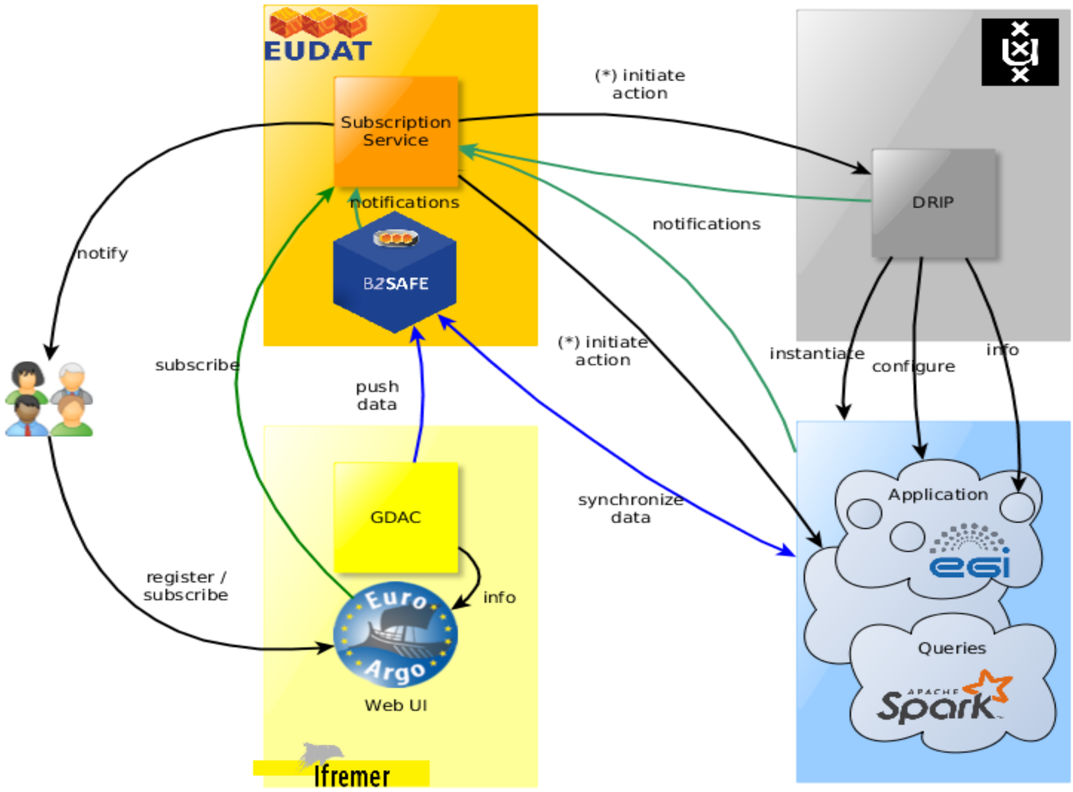

FIGURE 5. SYSTEM ARCHITECTURE FOR TEST CASE TC_2.

EGI and EUDAT explored the use of different technologies (Hadoop[6], OpenStack[7], Spark[8]) to store and query the provided dataset. Queries meet common marine scientist needs (all data from Mediterranean Sea, all temperature data from a specific platform).

---

[6] Hadoop: http://hadoop.apache.org/
[7] OpenStack: https://www.openstack.org/
[8] Spark: https://spark.apache.org/

A Web Portal has been developed to access indexed data. Data selection is made through different criteria (platform type, measure type, parameter, platform Id, time period and more). A python script has been provided by IFREMER to extract data satisfying a user's criteria.

EUDAT implemented a proof of concept of the data subscription service including a REST API. The service allows subscription to data though data identification. The service initiates associated actions and receives notifications. The use of the service and the API was demonstrated during the ENVRIweek in May 2017.

In this test case, the initiated action was the python script first synchronized from B2SAFE[5] to an EGI Federated Cloud[1] node and then initiated there with parameters. The process was executed using parallel computation provided by DRIP[9] (a service developed by WP7, T7.2), dynamically deploying and managing as many Virtual Machines as required to cope with the load in order to be able to process the subscriptions in a timely manner. Once results were available, they were pushed to B2SAFE[5] and a user was notified by email.

Next Steps

The next steps include a number of integrations (Figure 5):

1) GDAC-B2SAFE
2) GDAC-Web UI
3) Web UI-Subscription Service
4) Subscription Service-Apache Spark
5) Subscription Service-Cloud resources

The first integration will be realized through deployment of a suitable B2SAFE interface. Adoption of a common data identification schema precedes the second and the third integration. This also includes further refinement of the subscription REST API. The fourth allows queries as subscription actions to be executed on Apache Spark. In the fifth, the initiation of actions and provisioning of cloud resources is further refined.

Users should be notified by email (or SMS message) to inform the availability of result datasets in, e.g., B2SAFE[5] or B2DROP[10].

This development could be generic to a lot of scientific RIs. Diffuse Data, Synchronize them and Alert on update is a common needs for scientific data users.

3.3 Use Case TC_4: Sensor registry

Group Leader

Thomas Bloubrieu (IFREMER)

Objectives

The "sensor registry" aims at supporting the management of sensors deployed for in-situ measurements.

Four sub-use cases are considered:

- List and discover specifications of sensors and hardware on the market.
- Manage the park of sensor by owner, manage maintenance (e.g. calibrations), loan, …
- Edit deployments, enable traceability from observation data back to the sensor and procedures used for acquisition (link with implementation case on provenance IC_2).
- Discover infrastructures (observation network, equipped experiment sites...) and enable their citation.

[9] DRIP service is developed by WP7, T7.2: https://github.com/QCAPI-DRIP/
[10] EUDAT B2DROP service: https://www.eudat.eu/services/b2drop

The use case is applicable to the management of various types of platforms, deep-sea observatories (e.g. EMSO), marine gliders (e.g., EuroGOOS Gliders[11]) as well as solid earth (EPOS) or atmosphere (ICOS) observations.

This can also be used to track usage of specific sensor models (e.g. CO2) across the RI 's observation networks.

Partners

IFREMER, CU, Plocan, NERC.BODC, LOCEAN, IPGP, CNR/IBIMER, IPSL, 52°North, MARUM, CSEM, INRN

Related RIs

EMSO, EuroArgo, EPOS, ICOS, SIOS, ANAEE

Activities and Achievements

For the ENVRIplus Research Infrastructures, SIOS, EMSO, ANAEE, ICOS and CSEM, examples of sensor, site, station or platform descriptions have been collected in their native formats. The information models and tools used to manage the descriptions have also been shared.

Regarding standards, a sample of SIOS station description has been encoded in OGC/SensorML format. Examples of SSN (Semantic Sensor Network) ontology[12] encoding and translation from SensorML to SSN have been provided from FixO3 project inputs.

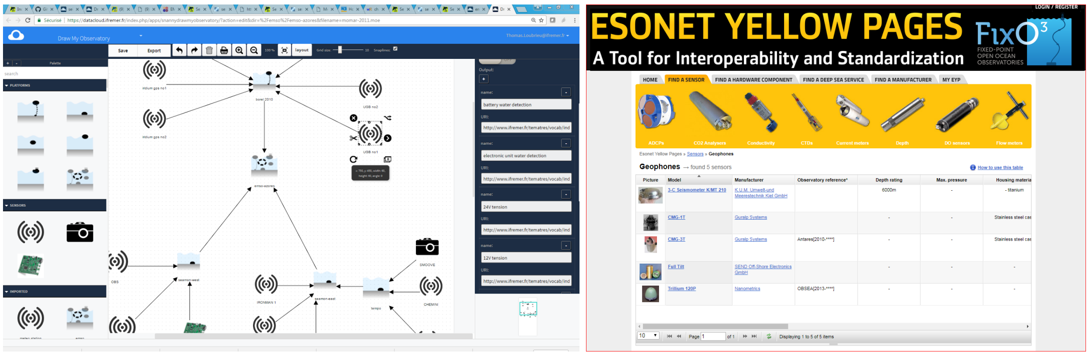

FIGURE 6. SENSOR DEPLOYMENT GRAPHICAL EDITOR AND SENSOR MODEL DATABASE FOR EMSO

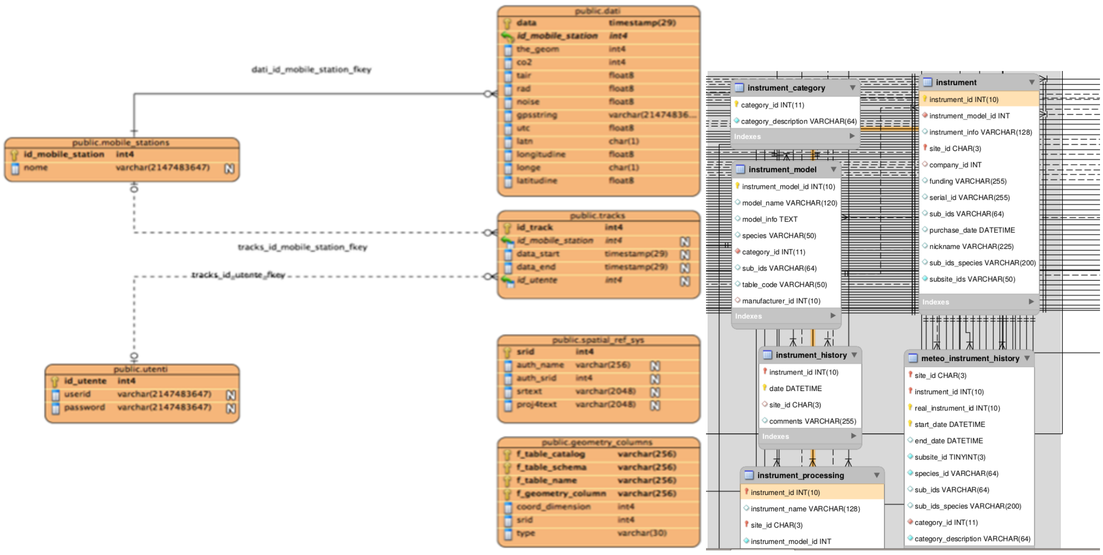

FIGURE 7. DATA MODEL FOR ANAEE AND ICOS SENSOR DESCRIPTIONS

[11] GuroGOOS Gliders: http://eurogoos.eu/gliders-task-team/
[12] SSN: https://www.w3.org/2005/Incubator/ssn/ssnx/ssn

Next Steps

The exploration of standards and tools to manage sensor, site, stations descriptions will be investigated further together the involved RIs and the experts in vocabulary standards (IFREMER, MARUM, BODC).

The descriptions managed by RIs in ad hoc format will be converted into standard formats (SensorML, SSN) and the integration of these descriptions in shared infrastructures based on these standards will be demonstrated by MARUM, BODC and IFREMER.

3.4 Use Case TC_16 Description of a national marine biodiversity data archive centre

Group Leaders

Dan Lear (MBA), Abreham Nieva & Alex Hardisty (CU)

Objectives

The Marine Biological Association (MBA) coordinates the DASSH Data Archive Centre, which is a national facility for the archival of marine species and habitat data. It needs to be integrated with other European marine biological data (e.g. data curated by EMSO, SeaDataNet, JERICO and EMBRC) as a joint contribution to EMODnet Biology, the COPERNICUS provider. This is a typical test case and will help showcase the benefits of using the ENVRI Reference Model (RM)[4] in a practical use case.

This use case will enable a full description of the DASSH Data Centre, and its associated services. It will support the integration with the EU data infrastructure, and serve as a demonstration of how the ENVRI RM could benefit other data centres.

The MBA staff involved has no prior experience with the ENVRI RM and are not ICT professionals. The implementation of the ENVRI RM by domain-centric experts could highlight differences or uncertainty in definitions and terminology.

We would like to describe the entire data flow from acquisition to publication. Existing flow diagrams are included below, with both internal and external flows and processes illustrated.

Partners

MBA, CU, UvA

Related RIs

DASSH, MEDIN

Activities

CU, MBA, and UvA have been collaborating on the use of the ENVRI RM to model DASSH. The initial tasks included basic training in the use of the ENVRI RM, followed by continued support and consulting during the modelling of different aspects of the RI. Table 4 summarises the main events and the outcomes/products associated with them.

TABLE 4. SUMMARY OF ACTIVITIES FOR TEST CASE TC_16

Activity	Partner(s)	Period	Result
Definition of DASSH use case	MBA	January 2016	Outline of the use case
Introduction to the use of the ENVRI RM	CU, UvA, MBA	February 2016	Basic RM modelling guidelines
Modelling and initial appraisal of use case (Second ENVRI Week Zaandvort)	CU, MBA	May 2016	Presentation of the DASSH use case initial analysis.

Independent Modelling of DASSH, from the Science and Information Viewpoint perspectives	MBA	May-July 2016	Initial set of DASSH models, hand drawn. Need for a tool for sharing models.
Proposal of UMLet and Templates	CU	July 2017	Simple open source UML editor tool to support modelling. A set of modifiable templates.
Modelling of DASSH using the UMLet templates	MBA	August-September 2017	Set of UML SV and IV diagrams of DASSH.
Report of advance and lessons learned (Third ENVRI week, Prague)	MBA, CU	November 2017	Report of advance, lessons learned, and pending work
Design of course contents based on the DASSH use case	CU, MBA	November 2016 until now	Curriculum for basic training on using the ENVRI RM.
Refinement of Curation and Publishing process models	MBA	January 2017 until now	Ongoing refinement of process models

Achievements

- An example of the detailed modelling of an RI and its needs for further development using the ENVRI RM.
- A set of models and model templates for the Science, Information, and Computational Viewpoints.
- A set of training materials for using the ENVRI RM based on a real use case.

Evolution of Data Flow Diagrams

Figures 8 and 9 show initial diagrams that MBA used to describe the DASSH data flows from acquisition to publishing. Figure 8 explains the detailed processing of new data assets as they are received, curated, and published. Figure 9 illustrates a higher view which focuses on the institutions that participate in the process.

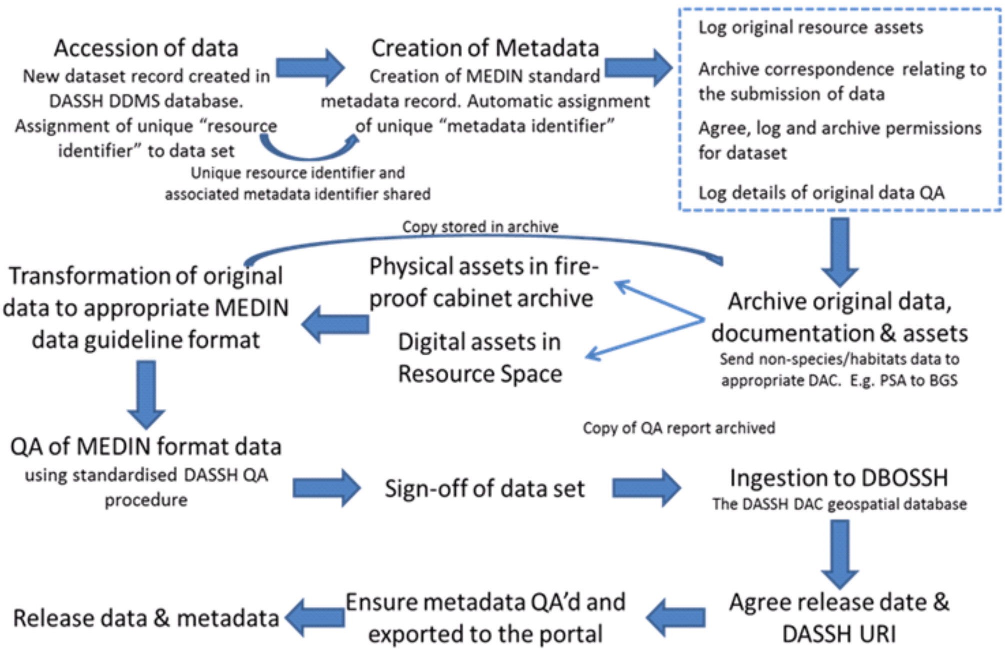

FIGURE 8: DASSH DATA MANAGEMENT PROCESS

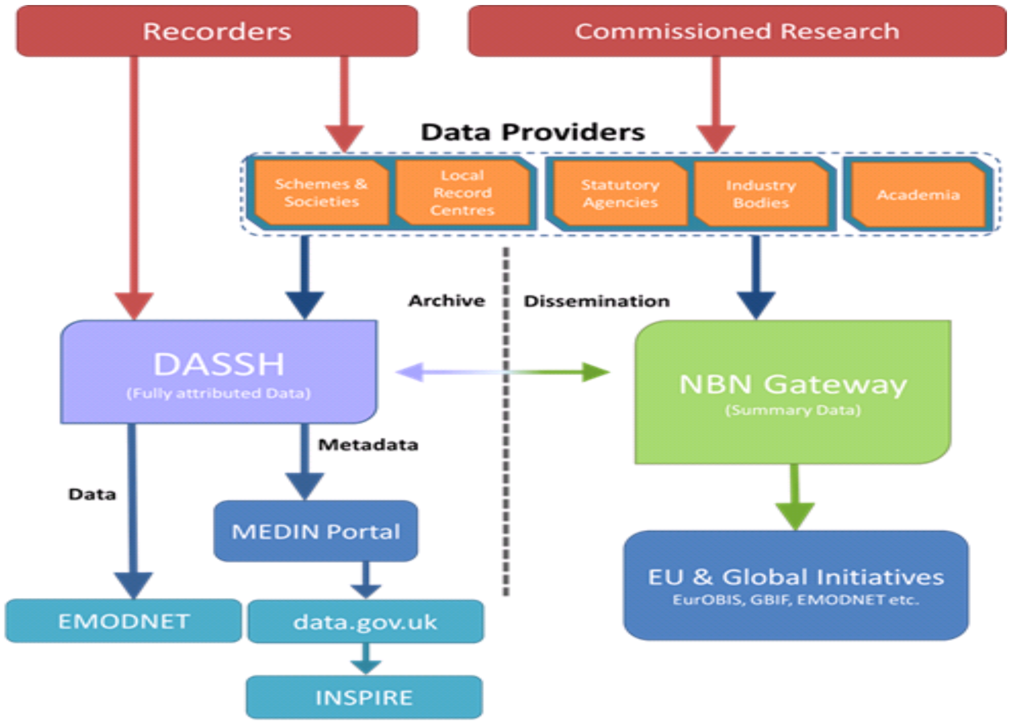

FIGURE 9: DASSH DATA FLOW

The diagrams shown initially contain useful information that helps understanding the data processes within DASSH. However they mix several concepts such as institutions, roles, processes, activities, information and information states. They are also acronym heavy. As a result, they require lengthy descriptions, which need to be adapted/changed depending on the public to which they are shown. As an alternative, DASSH has produced a set of UML diagrams following the ENVRI RM guidelines. The latest examples of these diagrams are shown in Figure 10 and 11.

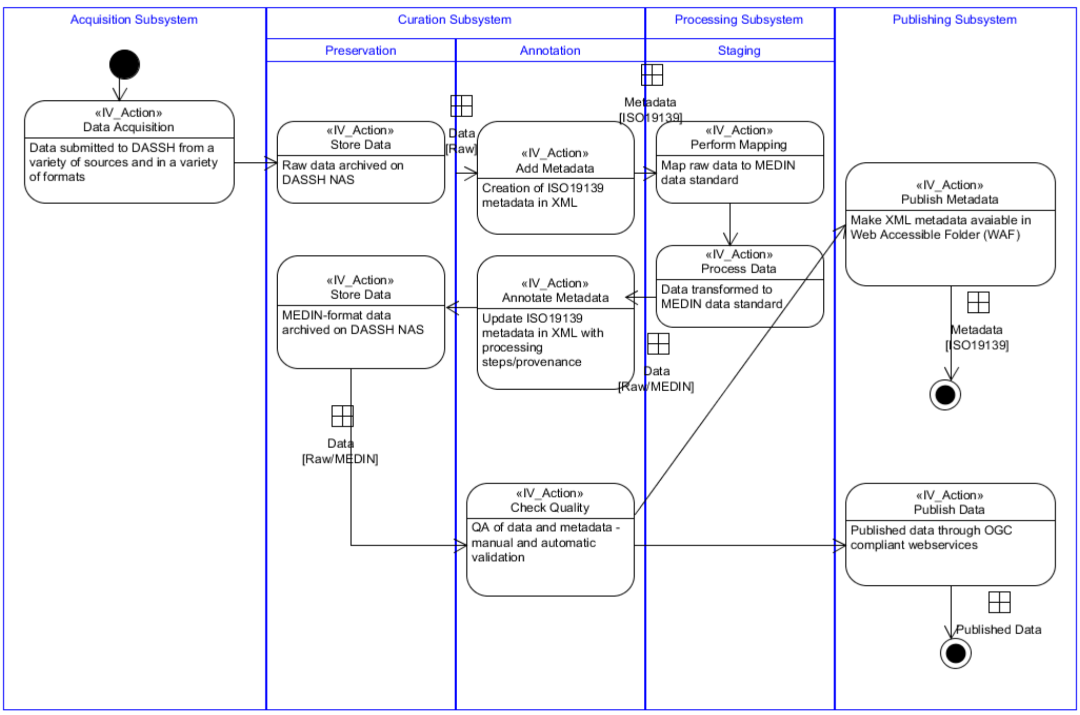

FIGURE 10: SCIENCE VIEWPOINT DIAGRAM OF DASSH DATA FLOW.

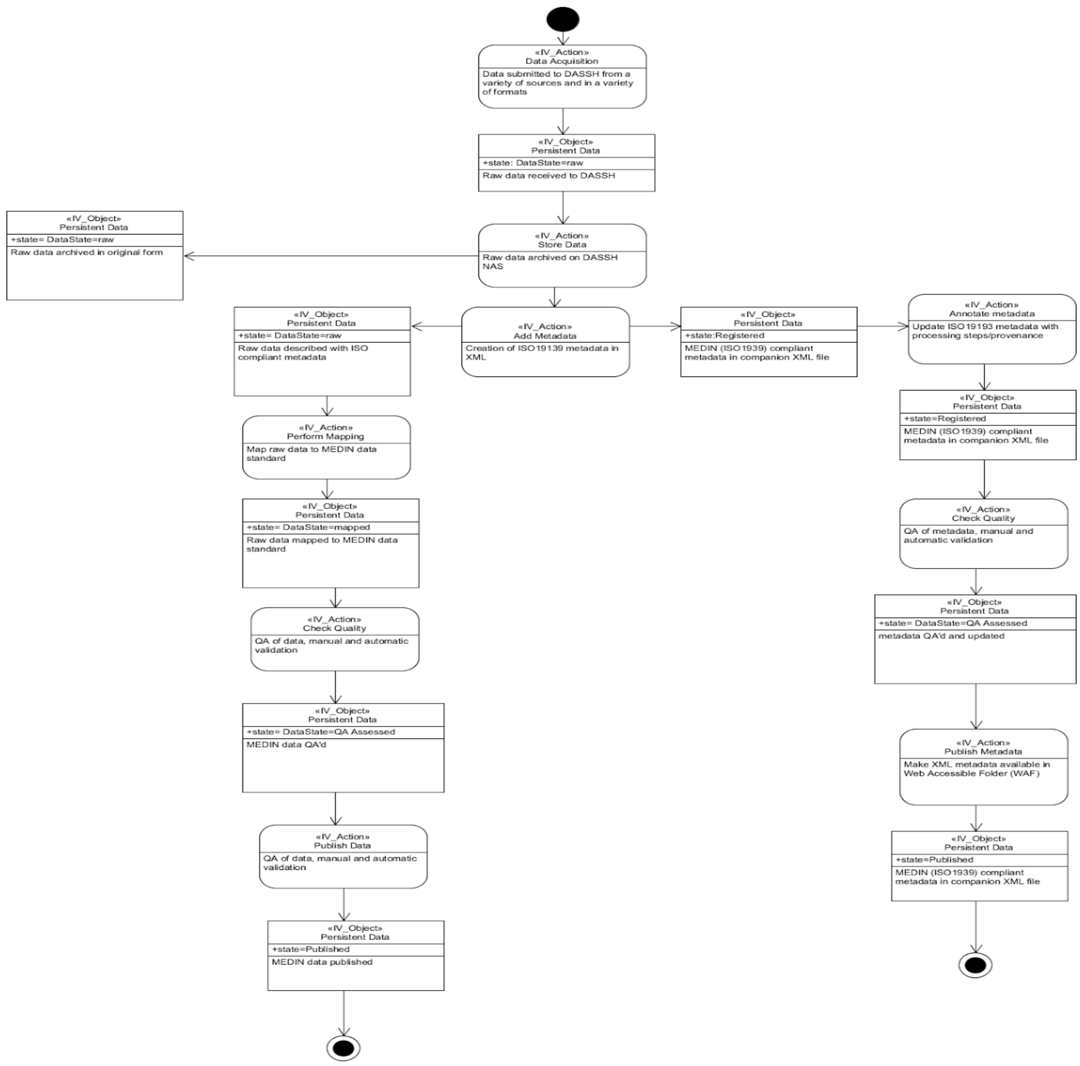

FIGURE 11: INFORMATION VIEWPOINT DIAGRAM OF DASSH DATA FLOW (FRAGMENT)

Using the ENVRI RM, MBA was able to refine the diagrams to illustrate different aspects of the processes. In Figure 10, the diagram emphasises the processes (actions) and the roles which perform those processes (systems and subsystems represented as pools and swimlanes). In Figure 11, the diagram emphasises the changes in data (data states) and the actions that cause those changes. These diagrams show clearly the correspondences between the science viewpoint and the information viewpoint. The Science Viewpoint diagrams are suited for presenting high-level details to all stakeholders. The Information Viewpoint diagrams provide further details which could be more useful for technical stakeholders.

Next Steps

There is a need for sharing of the model templates and examples, linked to RM definitions. Until now, DASSH and CU have been using UMLet based templates because they are encoded as XML and have the potential of being mapped to ENVRI RM constructs using the OIL-E ontology[13] for verification of correspondences. This will be further developed within Task 5.3.

[13] OIL-E: http://www.oil-e.net/

3.5 Use Case IC_1: Dynamic data identification & citation

Group Leaders

Alex Vermeulen & Margareta Hellström (LU)

Objectives

The objective of this implementation case is to define the best candidates for technologies that will allow implementation of data citation for dynamic datasets and collections of datasets. (Here, "dynamic data" refers to datasets that may change over time, e.g. because of new data points being appended, updates and/or error corrections, or changes in storage technology requiring reformatting.)

The work will focus on identifying methods following the recommendations of the RDA Data Citation Working Group (RDA DCWG) for a query-centric approach towards retrieval and subsequent citation of dynamic data [Rauber 2015, Rauber 2016]. This approach "allows us to identify and cite arbitrary views of data, from a single record to an entire data set in a precise, machine-actionable manner" by "identifying data sets by storing and assigning persistent identifiers (PIDs) to timestamped queries that can be re-executed against the timestamped data store" [Rauber 2015].

The expected outcomes – working demonstrators that cover two approaches (see Figure 12) for data storage: "flat files + catalogue database" and "versionable database containing both metadata and data" – will be submitted to the RDA DCWG as Adoption Use cases.

IC_1 is tightly connected to the ENVRIplus "pillar" Data Identification & Citation, and the expected outcomes are either dependent on, or should be interoperable with, several of the services under development within WP6. There are also strong links with two other "pillars", namely Cataloguing and Provenance, and thus care must be taken to ensure that the metadata schema utilised by the IC_1 demonstrators is aligned with the recommendations, as well as with e.g. metadata cataloguing services, arising from WP8.

Regarding assessment of the relevance and usability of the IC_1 outcomes, this is expected to be achieved primarily through collecting experiences from RIs that consider the implementation of the demonstrators into their own data curation.

Partners

ICOS (lead), AnaEE, IAGOS, ACTRIS, UniHB/PANGAEA, EAA

Related RIs

ICOS, AnaEE, IAGOs, ACTRIS, EMSO, LTER-Europe

Activities

As part of the WP6-defined "use case study 2" (see DoA), ICOS has developed its ontology-based cataloguing database to enable the storage of pertinent data object (DO) metadata, including not only the DO creation date, PID and the date range of the contained values, but also version number and history-relevant information such as "supersedes"/"is replaced by" with pointers to other relevant DOs. The metadata concept is based on storing all entries (and deletions) as time-stamped assertions in RDF triple format (implemented in a Postgres data table as triplets with ISO formatted timestamps for each record). This approach allows traversing the catalogue in time, extracting snapshots of DO properties at any given time. This functionality is critical in order to support the correct resolution of stored search queries (which constitute the citable entities in the RDA DCWG approach) into the relevant data objects (flat files) or extracts/subsets (from versionable databases) *as they existed at the time of the original query* (or indeed at any other point in time).

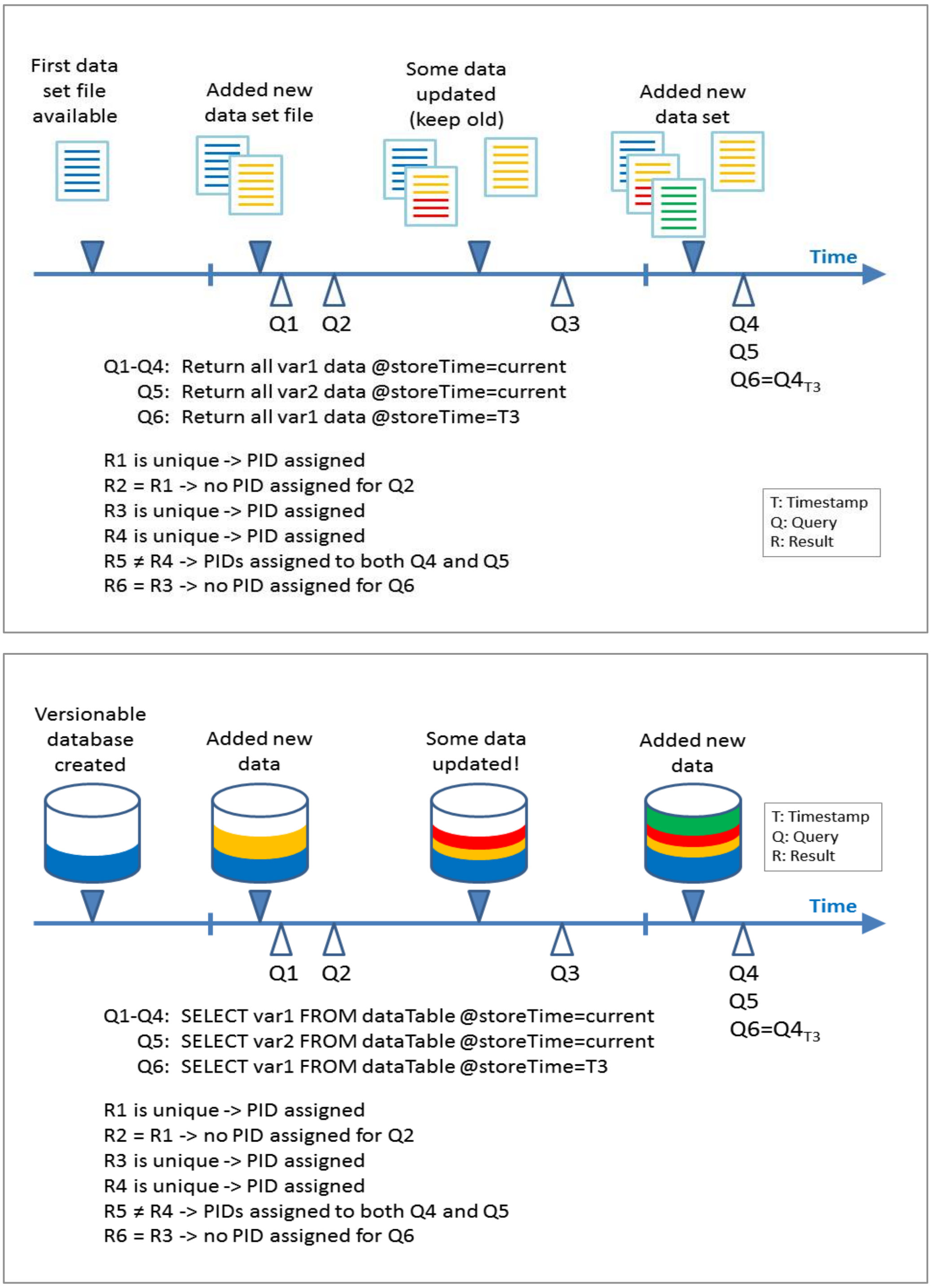

FIGURE 12. ILLUSTRATION OF HOW THE CHOICE OF DATA & METADATA STORAGE STRATEGY INFLUENCES QUERY-BASED DATA SEARCH AND RETRIEVAL. UPER PANEL: "FLAT FILE" OPTION; LOWER PANEL: "VERSIONABLE DATABASE" OPTION.

As a complement to the combination of a dynamic metadata catalogue with a traditional file system, ICOS has also investigated various options for “versionable” database technologies that may be used to store data values, together with timestamps identifying when every individual record was stored and when its status changed. This allows for “deletion” or “update” operations to be carried out without ever erasing records, and again permits the recall of the database status at any given point in time.

Achievements

Figure 12 shows some example queries and their results for a specific scenario in which data change over time. It illustrates that it is possible to achieve the same functionality using two different architectures.

In the first case, datasets are stored as distinct files, each with its own set of catalogued metadata - including file name, file creation date, date range covered by the dataset contents, and dataset version indicator. Following a specific dataset, for example observations of meteorological parameters from a station ìAî, we see first one data file (blue; from e.g. January of 2016) is uploaded, followed by a second dataset (yellow; e.g. February 2016). At a later time, an updated version of the February 2016 dataset is made available (yellow-red). However, the original February data file is still available. Finally, after some more time has passed, a third dataset from ìAî, e.g March 2016, is uploaded (green).

The second case shows the analogue situation, when a versionable database is used to store the observational datasets. Note that in the case of the modified February data, both the original and the updated data values are kept in the data store.

If an end user queries the data repository, the processing mechanism will use the information on in the catalogue to decide on what to return. The Figure shows six Queries performed at four different points in time, and the respective Results. Only results that are unique, based on their actual content, will be assigned PIDs, which can then be used in citations.

Both alternatives will require completely different implementations and the choices made will have considerable effects on performance and scalability. We expect that the flat file based approach is slightly more complex but will be more flexible, portable and performant.

Next Steps

Work will continue during the coming months, including both a follow-up of RDA activities related to the original DCWG work, and the preparation of the demonstrator systems (based on ICOS).

However, as pointed out during the initial evaluation of the IC_1 proposal, technologically successful solutions alone are not a guarantee for enthusiastic or comprehensive uptake of a service or its underlying methodology. To increase the usability of dynamic data-related applications, and make the approach of IC_1 more generic to support other RIs than the ones directly involved, additional functionality – such as the automatic creation of proper attribution text for subsets and versions – should be developed in parallel to the work on the data and metadata catalogues and query stores.

3.6 Use Case IC_2: Provenance implementation

Group Leader

Barbara Magagna (EAA)

Objectives

This implementation case is to validate service results of Task 8.3, inter-RI data provenance and tracing services. In close cooperation with EUDAT, a cross RI provenance model is proposed to be developed as an extension of the EUDAT B2PROV service[14]. This model will be used by workflow

[14] B2PROV is a new EUDAT in planning service

ENVRI plus

engines (such as Taverna, dispel, GEF), and EUDAT services such as B2HANDLE[15] (service for persistent identifiers), B2NOTE[16] (service for data and metadata annotation), B2SHARE[5] (data repository service together with a service for persistent identifiers and a service for metadata). The aim of the provenance model is to track lineage through the data lifecycle, and to support the data users when they access this information. It helps to identify datasets that are derived from the same source but displayed, published or processed in different ways or by different data services. It also helps to acknowledge any processes in the data lifecycle and the actors corresponding to these processes. It supports data citation, which will help encourage researchers to publish their data.

Partners

EAA, EUDAT, UvA, INRA, DLRZ, CNRS, INGV, LU, CINECA

Related RIs

LTER-Europe, ANAEE, IS-ENES, EPOS, ICOS, ACTRIS

Activities

Since Task 8.3 was only started in May 2017, the implementation of the case has not started yet. However, the agile team has made the following plans for the work:

1. Define minimum information that has to be tracked
2. Find a conceptual model for provenance that conforms to the needed information
3. Map existing models to the common model
4. Find a repository to store the provenance information
5. Design and implement provenance service
6. Design and implement publication service that allows provenance information to be delivered together with data and metadata

We will start with examining existing provenance implementations in ENVRIplus RIs and capturing the concepts and functionalities within the ENVRI Reference Model. Our timeline for the implementation is as follows:

- 5th ENVRIweek (November 2017): official start of activities
- 6th ENVRIweek (May 2018): presentation of first results
- 7th ENVRIweek (November 2018): final demonstrations

Achievements

Although the implementation case has not started officially, some preparation steps have already been carried out with preliminary results:

- Developed UML module for provenance (PROV[17]) using Enterprise Architect (EA)
- First draft of provenance representations of data collection in LTER-Europe in EA (see Figure 13)
- Started collecting existing provenance solutions using EA, including:
  - EPOS solution which uses dispel4py workflow engine in VERCE
  - IS-ENES approach within CMIP6, climate models
  - D4science provenance functionality
- Drafted a first concrete use case: combine PIDservice (e.g. B2SHARE) and ORCIDservice with annotation service (B2NOTE) linked to EnvThes to describe LTER dataset provenance related information

---

[15] B2HANDLE: https://www.eudat.eu/services/userdoc/b2handle
[16] B2NOTE: https://www.eudat.eu/tags/b2note
[17] PROV: http://www.w3.org/TR/prov-o/

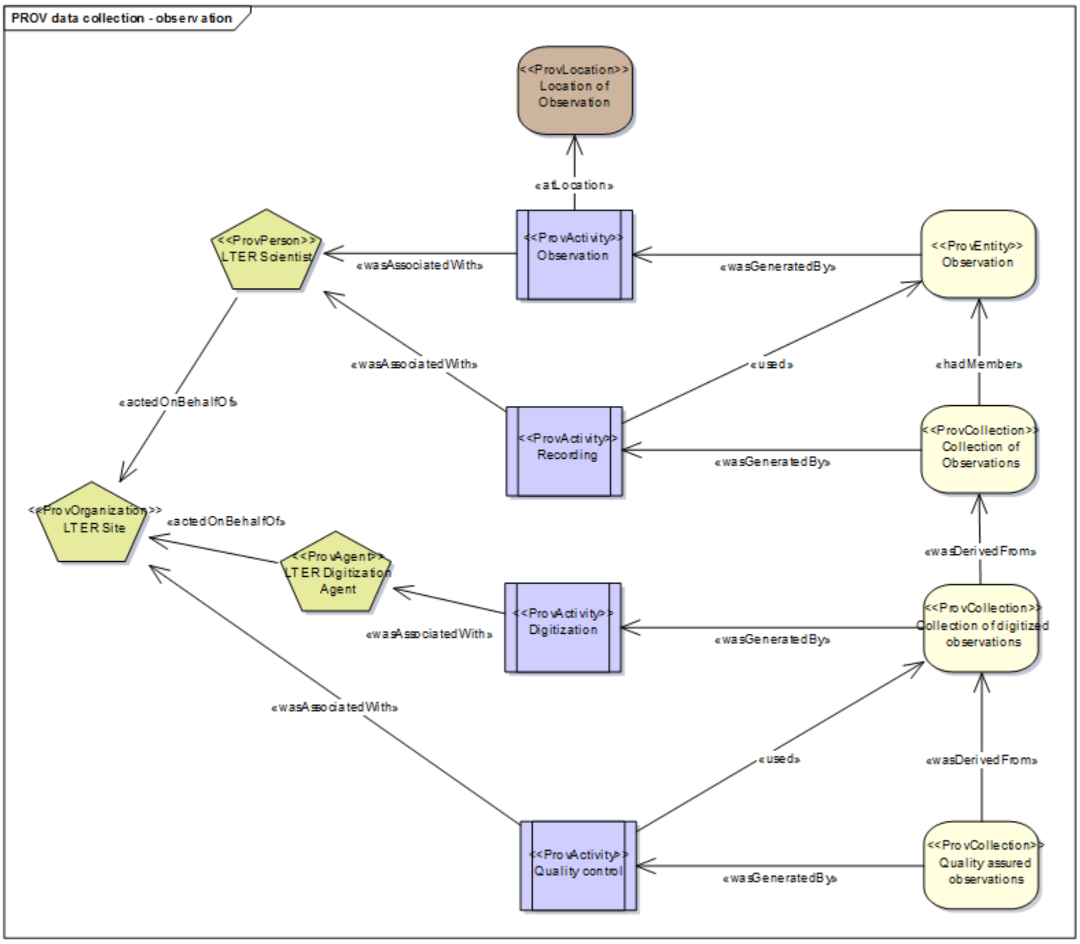

FIGURE 13. PROVENANCE MAPPING OF LTER-EUROPE DATA COLLECTION IN ENTERPRISE ARCHITECT

Next Steps

We will follow the defined work plan to achieve the investigation goal of this use case. Immediate steps to continue the current activities include:

- Identify missing elements in ENVRI RM viewpoints and extend the RM to support the description of provenance service implementations in a RI
- Finalize UML module for PROV
- Finalize UML representation of LTER-Europe data collection
- Analyse dispel4py workflow engine (EPOS solution)
- Evaluate first possible use case – combination of PIDservice/ORCIDservice/B2NOTE for LTER-Europe
- Prepare concrete implementation case draft for 5th ENVRIweek

3.7 Use Case IC_3: Support EISCAT_3D users to reprocess data using user's algorithms

Group Leaders

Ingemar Häggström (EISCAT), Leonardo Candela (ISTI-CNR)

Objectives

EISCAT_3D is an environmental research infrastructure on the ESFRI (European Strategy Forum on Research Infrastructures) roadmap. Once assembled, it will be a world-leading international research infrastructure to study the atmosphere in the Fenno-Scandinavian Arctic and to investigate how the Earth's atmosphere is coupled to space.

In general, EISCAT data products (including future EISCAT_3D data) are not raw data as sampled in the receivers, but have already undergone several steps of processing (along the chain voltages → filtering → time averaged spectral data -> inversion of physical parameters) by standard analysis algorithms in the EISCAT ICT system. This also implies that data are processed with a predefined set of parameters, e.g. spatial and temporal resolution, inversion model selection, model parameters and allowed parameter ranges.

This use case addressed a requirement of the EISCAT user community, namely to allow individual scientists to process their experimental data using their own algorithms. The challenge in this use case is how to make use of ENVRIplus services to demonstrate that EISCAT scientists can re-process data by defining other selections of parameters and algorithms.

Partners

EISCAT and ISTI-CNR

Related RI

EISCAT_3D

Activities

A typical usage scenario is that users select data together with an algorithm and invoke it with tuned parameters. The data to be used in the test case is from the present EISCAT facilities, and the processing software is provided by EISCAT and originally written in Matlab. In this use case, we chose a Matlab processing package that could be converted into open source software. The process is to generate a graphical visualisation of the experimental data. Figure 14 shows the EISCAT Real Time Graph (RTG), which is primarily developed to follow the radar run in realtime and to produce a display of basic parameters, such as returned power spectra.

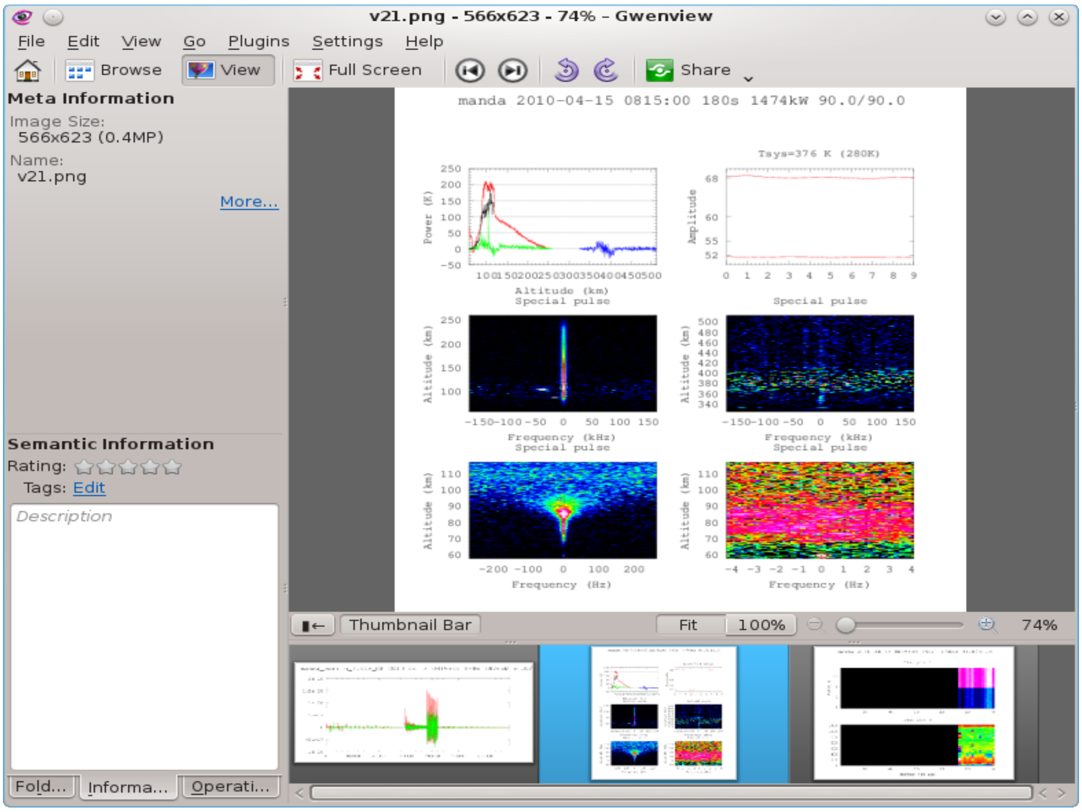

FIGURE 14. EISCAT REAL TIME GRAPH PRODUCES VISUALISATION OF THE PROCESSING OF RAW EISCAT RADAR DATA WITH BASIC PARAMETERS.

Achievements

The RTG was modified to run under GNU Octave[18], and to produce the plots in batch mode with some selected input parameters. The whole setup was installed into the gCube Virtual Research Environment, provided by ENVRIplus WP7. The gCube Data Analytics is executed via a web dashboard based GUI. The ticketing system of d4science has been effectively used for tracking the implementation and communicating with WP7 developers. Some initial results are shown in Figure 15, which demonstrates that in the gCube platform, the EISCAT RTG algorithms are translated into Octave, and compiled and published in ENVRIplus community space. In the same way, an algorithm created by a single EISCAT user can be shared and reused by others.

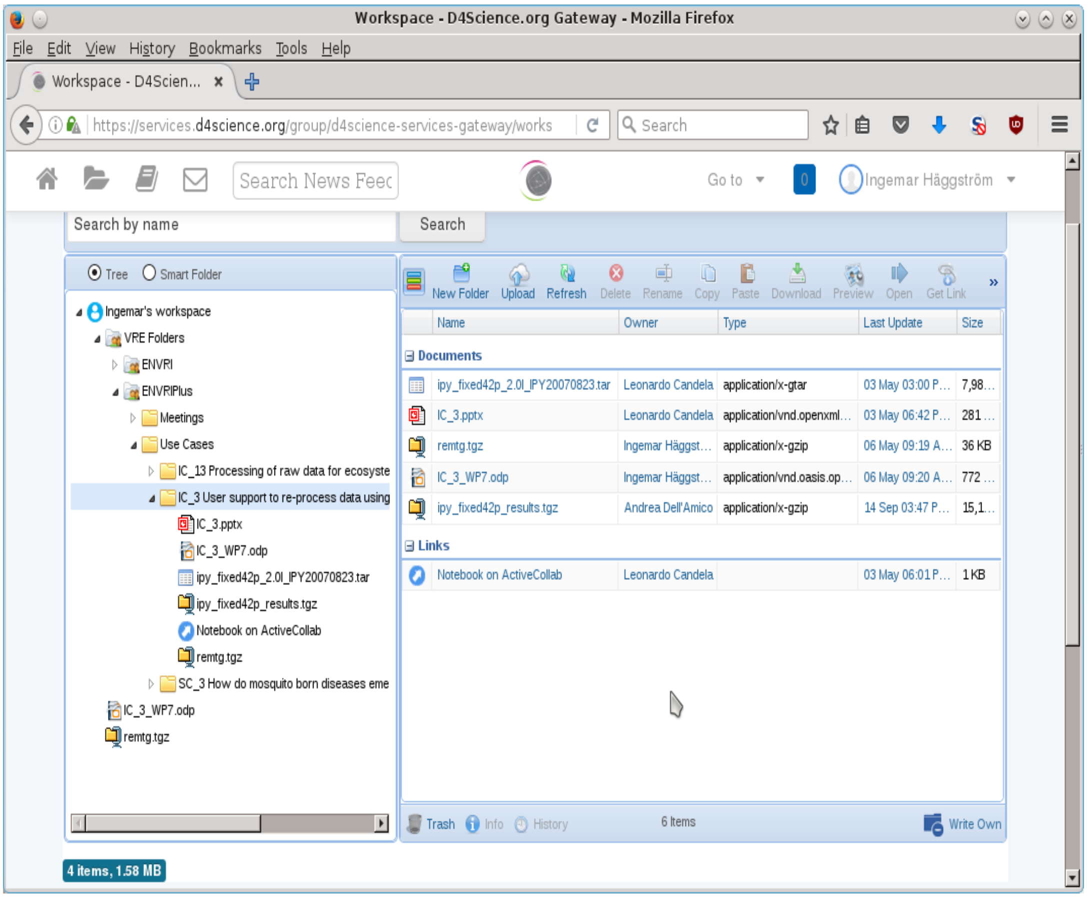

FIGURE 15. IN GCUBE, EISCAT USERS CAN CREATE PROCESSING ALGORITHMS AND SHARE WITH OTHERS.

Next Steps

The process needs to be completed with an improved graphical interface. At the moment, the pictures produced turned out to be somewhat difficult to grasp, so the software has been developed to compile them to .webm movies. This has not been fully implemented into gCube. Another issue is that the data is not under federated storage, and the new interface should use the storage archive at EISCAT Headquarters, instead of the test data at gCube. For this purpose it is necessary to consider how access to the data should be solved. A DIRAC[19] Storage Element is one possibility to stage data from EISCAT (<1TB) at gCube during the process. The size of the output data is a few MB and this should not be a problem. When the output is ready, the user should be alerted.

---

[18] GNU Octave: https://www.gnu.org/software/octave/
[19] DIRAC: http://diracgrid.org/

3.8 Use Case IC_9: Quantitative accounting of Open Data use

Group Leader

Markus Fiebig (NILU)

Objectives

In order to maximize the benefit of public investments into collection of geoscience data, funding agencies are pushing towards re-use of data for multiple purposes, including re-distribution and commercial use. The re-use of data for other than the original purposes is supposed to facilitate new services, thus generating economic growth. This vision requires increasingly more open data policies. On the other hand, many of these observations are collected by scientifically oriented RIs, where documentation of scientific merit in form of citations or use is paramount for scientists' employment and stations' funding. The objective of this use case is thus to **facilitate quantitatively correct accounting of data use in an open data world.**

This implementation case relates to, and is of cross-cutting relevance for, Task 6.1 of WP6. This task deals with various aspects of data identification and citation services across RI domains. Distributed over various actions, this includes citation tracking with respective interaction with publishers, and publication of respective best-practice documents.

To implement this use case's objective, some functionality needs to be in place:

Functionality 1 Primary identification of **all** data archived in a data centre with suitable but homogeneous granularity. The granularity needs to be fine enough to resolve authorship of a dataset down to an individual principal investigator (PI).

Functionality 2 Identified data collections need to include references to the primary identifiers of all data contained in the collection.

Functionality 3 Any services quantifying data use need to refer to the primary identifiers issued by the data's primary data centre as accounting reference. References to identified data collections need to be resolved to the primary identifiers to facilitate correct quantification of data use resolving the author / principal investigator's contribution.

This use case works specifying Functionality 1 and 2, and works towards initiatives implementing Functionality 3.

Partners

NILU, ICOS, ETHZ

Related RIs

ACTRIS, ICOS, EPOS

Activities

The following activities are on-going to achieve the objectives:

1. The community of atmospheric RIs in ENVRIplus, i.e. ACTRIS, IAGOS, and ICOS, have been identified as a reference group. Using its interoperability meetings as a forum, the group has agreed to use IC_9's approach for data identification and to implement Functionality 1.
2. The Research Data Alliance (RDA) has recently established a working group on research data collections[20]. ICOS represents ENVRIplus and IC_9 in this working group, with the aim of including Functionality 2 in the specification currently being written by this working group.

[20] RDA Working Group on Research Data Collections: https://www.rd-alliance.org/groups/research-data-collections-wg.html

3. The most relevant European scientific indexing service focussing on observation data is DataCite[21]. Consultations with DataCite by ENVRIplus partners are on-going, e.g. through the Technical and Human Infrastructure for Open Research (THOR) project[22]. This interaction is utilised to investigate options for implementing Functionality 3.

Achievements

The results of the agile investigation up-to-this-stage are as follows:

1. Each of the ENVRIplus RI's in the atmospheric domain is currently developing a work plan for implementing Functionality 1, or is already in the process of implementing it.
2. The work for lobbying in the RDA working group on research data collections for including in its specification a requirement for referencing any primary identifiers of datasets contained in a data collection in the data collection metadata is on-going.
3. While investigating options for implementing Functionality 3, difficulties have been encountered. When using the primary identifiers that resolve the contribution of individual PIs as reference for quantifying data use, it is implicitly assumed that the granularity of primary identifiers is homogeneous across scientific domains. However, granularity of primary data identifiers can vary widely across scientific domains. Especially life sciences use finer granularities for primary data identification than most other domains, which will lead to inflation of citation numbers when data collection identifiers are resolved to the primary identifiers.

Next Steps

For Functionality 1 and 2, the implementation work is being continued. Specification and results will be made available to the ENVRIplus community.

For Functionality 3, the scope of the objective needs to be revisited and redefined. It is a clear motivation for principal investigators to share their data openly when the use of their data can be quantified, including when the data is used as part of a data collection. While inhomogeneous granularity of primary data identifiers makes comparisons of data use impossible between scientific domains, comparisons of data use within a scientific domain will still be possible and useful. To this end, a secondary service provider, e.g. a research infrastructure, will be able to use primary indexing services to resolve data use inside data collections. Thus, quantitative accounting of data use will be a domain specific, on-demand service.

3.9 Use Case IC_10: Domain extension of existing thesauri

Group Leader

Barbara Magagna (EAA)

Objectives

The multilingual SKOS/RDF thesaurus EnvThes[23] integrates important terms used in long term ecological monitoring, research and experiments. It is based on the US LTER Controlled Vocabulary, and has been extended by QUDT[24] units and dimensions, EUNIS Habitats, INSPIRE spatial data themes and special concepts, needed by the EnvEurope and ExpeER communities. It links to other existing thesauri and vocabularies, such as GEMET, EARth, AgroVoc, EuroVoc and Wikipedia, and EEA vocabularies.

[21] DataCite: https://www.datacite.org/
[22] THOR project: https://project-thor.eu/
[23] EnvThes: http://vocabs.ceh.ac.uk/evn/tbl/envthes.evn
[24] QUDT: http://www.qudt.org/

In this implementation case, we want to extend EnvThes for the marine domain. Possible inputs can come from LifeWatch Italy, whose Interactions Thematic Centre developed a Phytoplankton Traits Thesaurus and is working on the Fish, Macrozoobenthos and Microzooplankton Traits Thesauri. Other RIs such as EMBRC could contribute to this extension as they are interested in linking with the biological community.

The overall objective is to advance the harmonization of the conceptual worlds of different RIs, thus enhancing semantic interoperability – a precondition for sharing metadata and data. Going forward, it is necessary to overcome the following challenges:

- Synonyms and exact matching terms: how to harmonize contradicting conceptual approaches in different communities.
- How to completely detect homonyms
- How to deal with contradicting views of different communities.

This use case investigates semantic harmonisation methods developed in Task 5.3.

Partners

EAA, Università di Salento

Related RIs

LTER-Europe, LifeWatch Italy, EMBRC

Activities

The main areas of agile activities are as follows:

Specify scope and implementation of EnvThes. Two work sessions have been organized by EAA in Vienna (March and November 2016) with the editor group of EnvThes to discuss and specify the specific scope, and for quality enhancement of EnvThes. EAA collaborates with the environmental informatics team at Centre for Ecology & Hydrology (CEH)[25] which works on linked data and semantic representation of metadata, specifically of observed properties. For the implementation additional internal workshops at EAA with the technical expert group of LTER-Europe were organized.

Enhance EnvThes quality to guarantee a good basis for harmonization. In order to provide a good basis for harmonizing different related thesauri, it is necessary to improve the quality of the EnvThes corpus. Following activities have been carried out:

- Check and correct errors:
  - duplicates
  - misspellings
  - plurals
- Check hierarchy (balance the hierarchy levels)
- Check consistency of content
- Check completeness:
  - definitions (by comparing with GEMET and Henderson's)
  - translations (just started in Greek, German, French and Italian, this is an on-going work)
  - links (has exact match)
- Resolve open issues:
  - abbreviations (-> alternative labels)
  - codes (-> alternative labels)

[25] Centre for Ecology & Hydrology (CEH): https://www.ceh.ac.uk/

Additionally, we directly extended EnvThes by using MS-Academic keywords to select only domain relevant concepts (almost 1000 out of the first 10.000 most used keywords in academic publications), and by adding measures and methods used in remote sensing by domain experts.

Extend the semantic description of observations and measurements by implementing a core ontology. We have compared different semantic models for describing observations and measurements in the ecological domain and decided to adopt a three-step approach for addressing this issue.

- EnvThes stand alone should just be used for keywords selection in the metadata description in DEIMS
- EnvThes plus simple ontology which enables parameter harmonization by linking the complex measures concepts to the atomic properties
- EnvThes plus an observation ontology including sampling and methods descriptions

We have participated in the LifeWatch workshop 'Ontology & Semantic Web for Research', held in Lecce in July 2017. During the session on ontology alignment, a use case study was identified and will be incorporated in a common publication. The use case aims to measure a phytoplankton property semantically represented using different observation models (including SERONTO, a model developed by EAA) and evaluating the pros and cons of each approach. This will help identifying the appropriate ontology to be used for the whole description of the measurement property.

Identify and analyze candidate vocabularies for extension. We have participated at three related workshops: 1) Thesauri & Semantics June 2016 in Lecce, 2) Marine Species Traits workshop, June 2016 in Ostend, and 3) Ontology and Semantic Web for Research workshop, July 2017 in Lecce. These workshops help us to get an overview of candidate vocabularies for extending EnvThes.

Study mapping and linking options for the extension of EnvThes. We have participated at RDA VSIG (vocabulary service interest group) bi-monthly teleconferences as well as at the RDA 9th Plenary Meeting in Barcelona, April 2017. As a collocated event, EAA has co-organized a EUDAT semantic working group workshop with the focus on discoverability and interoperability of semantic resources. During this workshop Barbara Magagna co-initialized the revival of the RDA VSIG (which has paused for 6 months), which now has the name VSSIG (Vocabulary and Semantic Service Interest Group). EAA is also involved in several semantic tasks in EUDAT, e.g. in the development of the data annotation service B2NOTE as well as in the development of an Ontology Look Up service for a cross-repository aggregation. All these activities helped to get an overview of possible means to discover and compare existing vocabularies and ontologies in the domain. One of the most promising options seems to be to register EnvThes and other related vocabularies in one single platform in order to take advantage of the search and mapping capabilities of such registries. As a result of these interactions, LifeWatch Italy seeks cooperation with LTER-Europe to build a new ontology and vocabulary library for the ecological domain called ECOPortal, which should be launched at the beginning of 2018.

Once this portal will be available, and all related vocabularies will be integrated, external as well as internal mappings between vocabularies will be easily performed. A decision guidance in this mapping exercise will be developed so that this experience will be reusable for other domains (this output should be ready by May 2018).

Achievements

In summary, we highlight the key outputs from above activities as follows:

1. We have defined the scopes of EnvThes as follows:
    - Select keywords for the annotation of datasets, sites and data products in DEIMS. The information is stored in the generic keyword field, as well as in specific fields (e.g. observed parameters or research topics for the research sites)
    - Select proper common terms for translating local parameter names (e.g. in a data table or SOS service) into common parameter names agreed in the (LTER) community.

Both scopes are dedicated to information discovery. Both consider datasets and sites as well as visualizations of parameters observed.

EnvThes can be used in the following areas:

- Keywords: EnvThes contains all terms used for the keyword selection. This might encompass all terms in EnvThes. All keywords should be available in multiple languages to support translation needs.
- Simple parameter name harmonisation. To harmonise parameter names, a complex concept of general parameter names = 'measures' and atomic terms describing this observation will be used (see Figure 16). Atomic terms (e.g. properties) can be used for the discovery of data. This approach uses the methodology proposed by CEH[25] to represent the observable property[26]. In EnvThes both concept types will be accommodated, the complex as well the atomic terms.

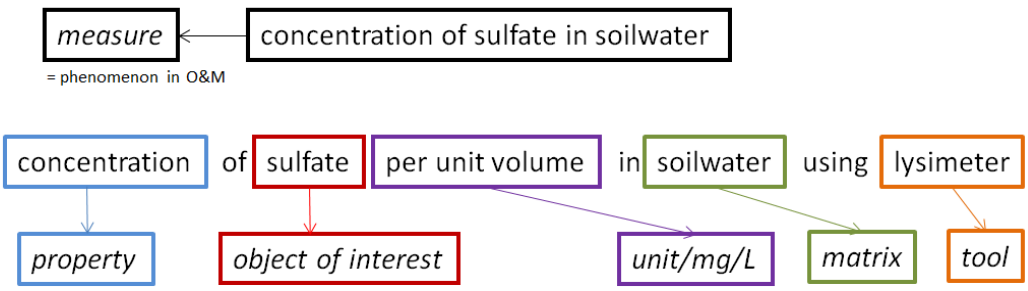

FIGURE 16. COMPLEX VERSUS ATOMIC CONCEPTS IN ENVTHES

Common parameter names are currently implemented in LTER-Europe. In the local data service the original parameter name (e.g. Lufttemp) and the assigned common EnvThes parameter name (e.g. Air temperature) are stored and exposed using, e.g., a Sensor Observation Service (SOS) endpoint (Figure 17). The Data Integration Portal (DIP) harvests the metadata on the server and exposes the term from EnvThes if it is mapped to EnvThes. Therefore, even if different local parameter names (e.g. language, abbreviation, coding) are used, a common representation of the data is possible. The discovery of the dataset parameter is then based on a full text search in the provided metadata and on the assigned common parameter name.

2. We have made good progress on quality enhancement of EnvThes. We have fixed a list of errors and duplications, resolved a number of open issues, and created extensions. Addition of definitions is in progress.
3. We have started to examine the state-of-the-art on observation ontologies, and comparison of different approaches (Lecce Workshop). We have defined a three-step approach outline for the LTER-Europe implementation.

---

[26] This is based on the SensorML 2.0 definition of an Observable Property in section 7.2.1 of the OGC specification, where it is only defined as something that can be observed, and that any units of measurement, quality information, or constraints (such as average, location etc.) should not be used, as these are artefacts of the observing process, not the property/phenomenon.

4. Vocabularies for extension of EnvThes have been identified, including the following products developed by ENVRIplus RIs:
    - The vocabularies developed by the Thematic Centres of LifeWatch Italy[27], which includes thesauri such as the Functional Traits Thesauri (Phytoplankton, Macrozoobenthos, Fish and Zooplankton as well as the Alien Species Thesauri
    - anaeeThes[28] of the AnaEE Infrastructure
    - SeaDataNet Vocabularies[29]
5. A publication, "EcoPortal: facilitating discovery and interoperability of ontologies and vocabularies in the ecological domain", is submitted to S4BioDiv Workshop in Vienna 2017, and currently under evaluation.

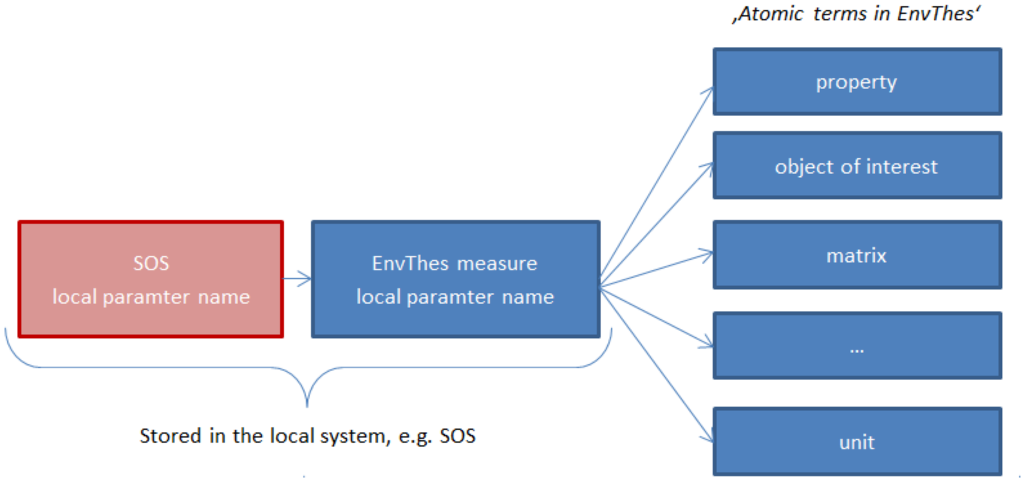

FIGURE 17. ILLUSTRATION OF HOW THE A SIMPLE PARAMETER HARMONISATION COULD BE ACHIEVED

Next Steps

We will continue the work on quality enhancement of EnvThes, and we plan to add definitions to concepts, add translations of concepts, and refine measures, objects of interest and properties according to the CEH approach on observable properties.

Towards the implementation of the core ontology, we will start developing the parameter harmonisation ontology based on the CEH property ontology, and decide on the observation model to be used as basis for the LTER ontology used to describe sampling and methods.

As the next step in analyzing existing vocabularies, we will identify the anchor concepts of the vocabularies developed by LifeWatch Italy, AnaEE, SeaDataNet, and map them to EnvThes concepts.

Additionally, we will help LifeWatch Italy to set up EcoPortal. This needs to define user community requirements and needed functionalities, identify quality criteria for including semantic resources, and test its services.

We aim to achieve the following by spring/summer 2018:

- Enable EnvThes internal mappings to vocabularies (exact match)
- Enable across thesauri (external mappings)
- Extend EnvThes with bridging concepts to enable linking between vocabularies
- Develop mapping guidance document

---

[27] Thesaurus Vocabularies developed by LifeWatch Italy: http://thesauri.lifewatchitaly.eu
[28] AnaEE Thesaurus: http://agroportal.lirmm.fr/ontologies/ANAEETHES?p=classes
[29] SeaDataNet Vocabularies: http://seadatanet.maris2.nl/v_bodc_vocab_v2/browse.asp

3.10 Use Case IC_11: Semantic linking framework

Group Leaders

Zhiming Zhao & Paul Martin (UvA), Barbara Magagna (EAA)

Objectives

In order to realise the convergence of vocabulary at a technical level, it is necessary to identify the shared concepts in various standards and specifications that currently govern the handling of data and tools in the environmental sciences. Open Information Linking for Environmental RIs (OIL-E) is a semantic linking framework being developed to help perform this challenging task based on the core concepts of the ENVRI Reference Model. In the context of this implementation case, we wish to construct a set of ontologies which we can use to describe the semantic landscape of RIs in ENVRIplus in a dedicated knowledge base. Within this knowledge base we can link directly to various published vocabularies, metadata and catalogues, as well as establish links with other ENVRIplus services such as the Flagship Catalogue Service[30]. We hope that this service will augment the activities of Task 5.3 of ENVRIplus (semantic linking), in particular by providing the persistent data store required for various different tools for creating and browsing semantic data required for that task.

Partners

UvA and EAA

Related RIs

This implementation case potentially involves all RIs in ENVRIplus.

Activities

The primary service being used is OIL-E[31], which provides a small set of ontologies based on the ENVRI Reference Model (RM) [32].

In the context of Task 5.3, a prototype knowledge base for RI data structured according to the OIL-E ontologies has been deployed and is being used as the basis for a number of experiments. For example:

1. A tool for expressing RI requirements and checking them against the RM is being developed.
2. The use of the RM for document annotation, specifically the identification of RM terms or synonymous concepts in the document text, is being prototyped.
3. Ways to link with the ENVRIplus Flagship Catalogue service are being explored.

The knowledge base is based on Semantic Web technologies, being implemented as an Apache Jena Fuseki[33] Web service providing a SPARQL endpoint for querying of the underlying RDF-compliant triple store. Based on the success of the prototype knowledge base, it will be proposed that a longer-term knowledge base be set up as part of the ENVRI community site/services. It is hoped that by using standard interfaces (SPARQL queries via a RESTful HTTP API, with responses encoded in JSON), the knowledge base can also be used by other ENVRIplus services for semantic data discovery or for process optimisation (for example using information

[30] Flagship Catalogue Service is described in ENVRIplus Deliverable D8.3 http://www.envriplus.eu/wp-content/uploads/2015/08/D8.3-Interoperable-cataloging-and-harmonization-for-environmental-RI-projects-system-design.pdf
[31] OIL-E: http://www.oil-e.net/
[32] ENVRI Reference Model: http://envri.eu/rm
[33] Apache Jena: https://jena.apache.org

about available resources and their operational requirements to help map applications workflows to infrastructure and network).

Achievements

An initial set of ontologies based on the ENVRI RM have been created, revised based on modelling experiences, and updated to maintain alignment with recent revisions to the RM. The OWL 2 ontologies are available at http://www.oil-e.net/ontology/ .

The current working version of the ontology model is split across three ontology files, compliant with the OWL 2 DL profile. More efficient (albeit less expressive) profiles are being investigated at the time of writing:

- **rm-core**[34] provides the base concepts needed to support the modelling of environmental research infrastructure.
- **rm-archetypes**[35] captures the set of objects, concepts and activities prescribed by the ENVRI RM for describing environmental RIs.
- **rm-correspondences**[36] provides additional information about correspondences between objects in the different viewpoints modelled by the ENVRI RM.

Within the OIL-E framework, it is now possible to develop a number of custom extensions covering different domains of discourse; it is also now possible to create mappings between the RM and different standards (e.g. CERIF) that might be used to describe specific objects prescribed by the RM.

Next Steps

Based on the prototype service currently deployed and the results of the experiments, it is hoped that a persistent knowledge base can be deployed as part of the ENVRI community portal[37], providing a programmatic map of the 'semantic landscape' of environmental science RIs. This semantic landscape will capture several aspects of the semantics-oriented activities being conducted currently be the RIs, in particular identifying the controlled vocabularies being used and linking to online (semantic) data sources. This could serve as an informational resource to the RI community and a record of ENVRI RM-oriented activities. In the meantime, it is intended to:

- Propagate prior descriptions of RIs in terms of the ENVRI RM into the knowledge base to serve as test data.
- Test the capabilities and limitations of the current SPARQL-based query interface into the knowledge base.
- Determine the suitability of the current ontologies for the kind of reasoning and information retrieval required by the various experimental applications described earlier and still to be conceived, and determine whether further refinements of the OIL-E framework are necessary.
- Prepare a demonstrator for the next ENVRIweek assembly in November 2017, as well as collect findings for the submission of D5.3 (regarding semantic linking in ENVRIplus) in April 2018.

[34] rm-core: http://www.oil-e.net/ontology/rm-core.owl
[35] rm-archetypes: http://www.oil-e.net/ontology/rm-archetypes.owl
[36] rm-correspondences: http://www.oil-e.net/ontology/rm-correspondences.owl
[37] ENVRI community portal: http://envri.eu/

3.11 Use Case IC_12: Implementation of ENVRI RM for EUFAR and LTER

Group Leader

Barbara Magagna (EAA)

Objectives

The objective of this implementation case is to describe the processes within EUFAR and LTER infrastructures using the ENVRI Reference Model (RM). The RM provides a set of ready-to-use terminologies with a public accessible reference base. This can be used to describe requirements and architectural features of the infrastructure, and serve as a common language in communication materials. Being a uniform framework with well-defined subsystems of components specified from different complementary viewpoints (Science, Information and Computation) the RM promotes structural thinking when constructing distributed and multi-organizational system architectures and can be used as a research tool for comparison and analysis of different technologies and solutions to guide design and implementation activities, and to drive the development of common services.

Partners

EAA, DLR

Related RIs

LTER-Europe, EUFAR

Activities

There are three groups of activities around this case:

EUFAR RM related activities. The expected output of the implementation use case for EUFAR will be a considerable help in the preparation of the research infrastructure in the process towards establishing a sustainable legal structure. RM allows roles and processes to be clearly defined in EUFAR's operations and helps designers describe the current situation as well as find missing or duplicated actions.

The first focus was the modelling of roles and behaviours in the science viewpoint. For the definition of behaviours, activity diagrams with roles, artifacts and steps were modelled. Following processes are represented:

- data collection (see below first part of the process model)
- data curation
- data discovery and access
- user behaviour tracking

Both the actual as well as the planned processes are in the process of being described.

It is not yet clear whether we will proceed with the representation in the information viewpoints. This would certainly include the correspondences relationships between science and information viewpoint. Roles and artifacts would correspond to information objects, artifacts in many cases states of information objects and action types would be transformed from steps. A workshop held in Vienna in March 2017 involving EUFAR helped to increase knowledge exchange and understanding in the use of RM modelling.

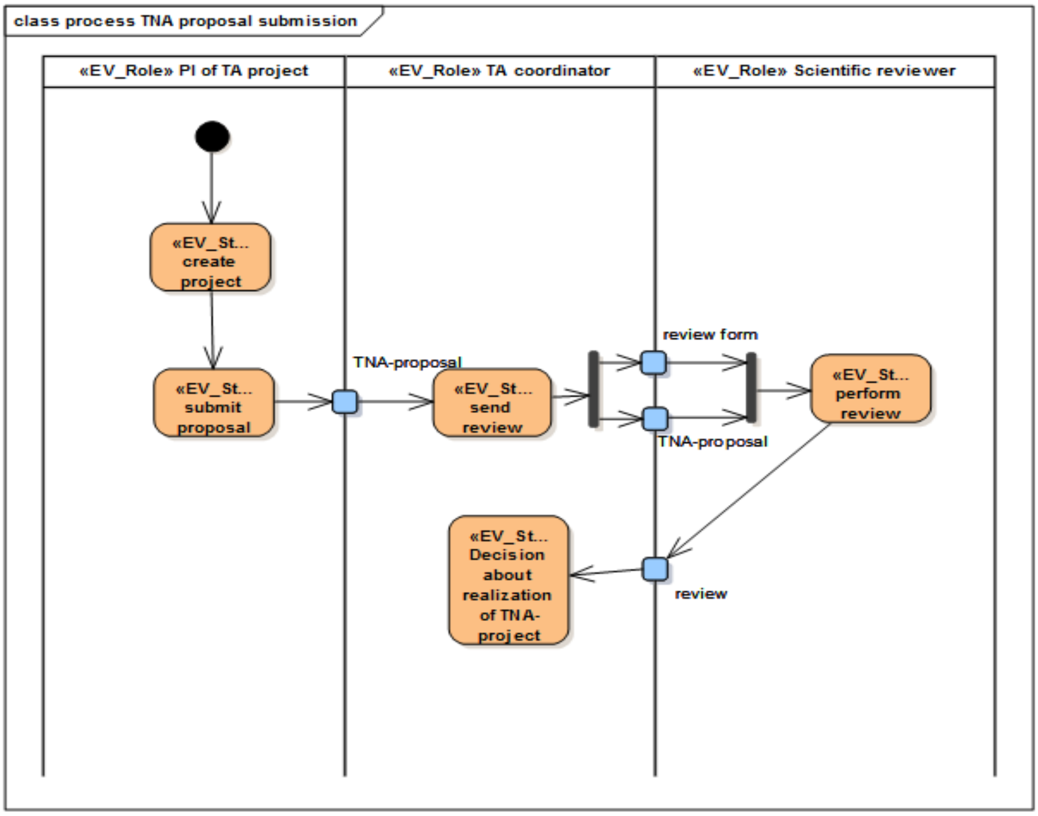

Figure 18. Activity diagram for TNA proposal[38] submission in EUFAR as a first process representation within the data collection behaviour model in Enterprise Architect

LTER RM related activities. LTER is a network of 500 LTER sites and 40 LTSER platforms. These sites have extremely heterogeneous data management capabilities and IT capacities:

- Basic capabilities: doing very basic data management (e.g. based on Excel spreadsheets), with no major IT support and infrastructure.
- Regular capabilities: providing well structured data storages (e.g. Databases) and related metadata. Ready to setup data services, but need support (e.g. tools and IT infrastructure) and guidance (e.g. which standards to use).
- Advanced capabilities: already implemented a comprehensive data management system with services installed (e.g. WFS, WMS, SOS, Linked Data)

The RM is used to represent the data architecture of LTER from the perspective of the Data Integration Portal team, taking care of the different approaches of the data nodes which will be of important help in the technical communication within the LTER user community. An RM representation of the science, information and computational viewpoints is being produced using the same procedure as in EUFAR, based on EA templates.

RM development related activities. The RM representation of EUFAR and LTER made evident that improvements/adaptions of the model's semantics and of its representations are needed.

- New semantic elements: role (not included in the RM ontology so far), person, institution, project, research infrastructure, objective, step, process
- New representations of RM: apart from the RM descriptions in the ENVRI wiki[39], UMLet templates were developed by Cardiff University (A. Nieva). Umlet[40] is free and easy to learn, but is only a graphic tool and thus lacks important features needed for effective

[38] EUFAR TNA programme: http://www.eufar.net/projects/ta-application/
[39] ENVRI RM wiki: https://confluence.egi.eu/display/EC/ENVRI+Reference+Model
[40] Umlet: http://www.umlet.com/

collaboration. EAA translated these models into Enterprise Architect (EA) diagrams, which can be used as templates with reusable elements. For EUFAR, EA is the better option as the licences are available at DLR and at EAA without additional costs. For better collaboration within a larger developing group it would be better to use a Subversion service but for only two people involved we decided to use Google Drive folders to save historical and actual versions.

For the very first steps of the modelling process EAA proposed a workflow where the RM expert prepares a Google Tables based template for first descriptions of processes to be filled by the RI expert. With these drafts it is easy to produce EA diagrams based on the templates (by the RM expert). These can easily be modified by the RI expert.

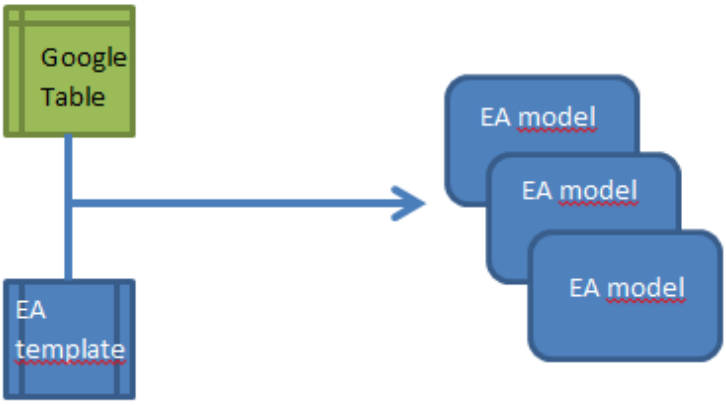

FIGURE 19. WORKFLOW OF MODELLING PROCESS (EA: ENTERPRISE ARCHITECT)

Achievements

According to the activity groups defined above, we highlight the following achievements so far:

1) EUFAR: Google Tables for all processes filled, EA models for data collection and data curation both actual and planned versions
2) LTER: EA model for data collection
3) EA templates for science, information and computational viewpoints

Next Steps

1) EUFAR:
   a) Finalize EA models for all processes
   b) Instantiate the RM ontology with EUFAR use case
   c) Write a report on the use of RM for EUFAR
2) LTER:
   a) Finalize science, information and computational viewpoints for all three types of nodes (basic, regular and advanced)
   b) Instantiate the LTER use case in RM ontology
   c) Write a report on the use of RM in LTER
3) RM representation:
   a) Finalize harmonization between RM wiki, UMLet and EA templates and RM ontologies
   b) Evaluate easy EA transformation to RDF to integrate those models in the RM ontology
   c) The use cases as instantiations of the ENVRI rm-archetypes will be incorporated into the ENVRI knowledge base

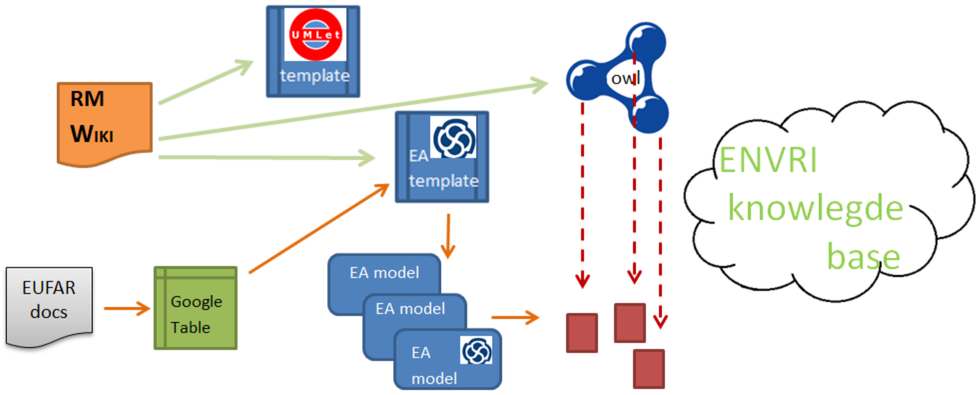

FIGURE 20. REPRESENTATION MEANS FOR MODELLING RI DRIVEN USE CASES WITH THE ENVRI RM

3.12 Use Case IC_13: The eddy covariance fluxes of GHGs

Group Leaders

Dario Papale & Domenico Vitale (UNITUS)

Objectives

The main objective of use case IC_13 is to optimize 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.

The EC fluxes are calculated using a complex set of data processing and quality control steps that require computational resources and knowledge of the technique. This is particularly important for the RIs (e.g. ICOS) that aim to process simultaneously multiple sites data and/or in Near Real Time (NRT) modality (i.e. provide each day fluxes measured the previous day).

The goal of the IC_13 use case activity involving experts in eddy covariance data processing and in computer science is to develop, test and provide a tool able to process EC data following a robust scheme, with uncertainty estimation, in an efficient way from the point of view of the computational time. This use case integrates gCube service (provided by WP7, Task 7.1) to deliver the desired functionalities. The approach is generic and can be used by any other RIs using EC data (e.g. LTER, ANAEE).

Partners

ICOS-ETC (University of Tuscia, Viterbo, Italy), ISTI-CNR

Related RI

ICOS

Activities

The ambitious goal is to provide an effective tool able to process any kind of EC raw data and offer to any user the possibility to calculate EC fluxes according to the highly-standardized EC processing scheme implemented by the ICOS RI. To achieve this target, it is straightforward to manage with care the EC metadata information (e.g. type of instrument used, EC system setup, file format and structure).

For ICOS RI, the metadata workflow is now established and a persistent digital identifier (PID, EUDAT services) to data object through the data processing lifecycle will be used for referencing data object in workflows and provenance records. In this way, raw data and their associated metadata will be linked together, thus facilitating data transfer and automated processing.

For other ENVRIplus RIs (e.g. LTER, ANAEE) that do not compliant with ICOS RI requirements, it will be required to standardize 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.

A simplified workflow representing the different way to manage raw data and metadata information is depicted in Figure 21.

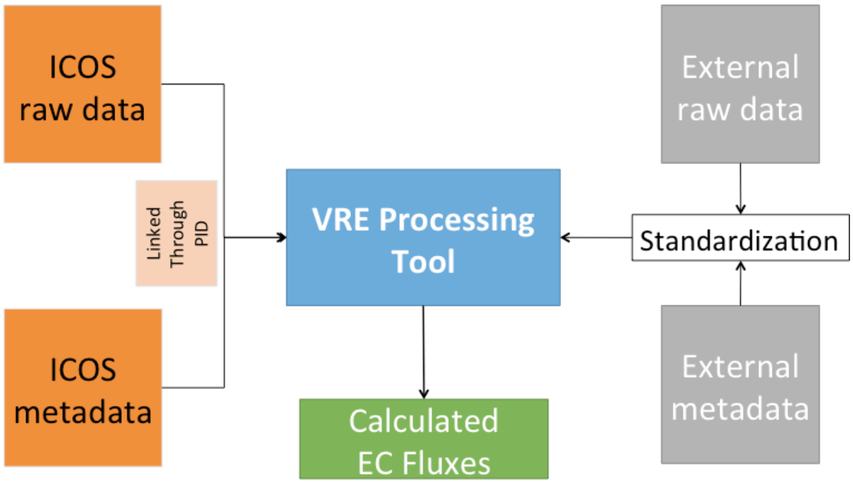

FIGURE 21. A SIMPLIFIED WORKFLOW OF RAW-DATA AND METADATA MANAGEMENT DURING THE PROCESSING STAGE.

Achievements

The EC data processing was enhanced to run under R environment with a call to EddyPro® Fortran code to process the high-frequency EC raw data. In agreement with ICOS protocol, the processing path is defined as in Figure 22.

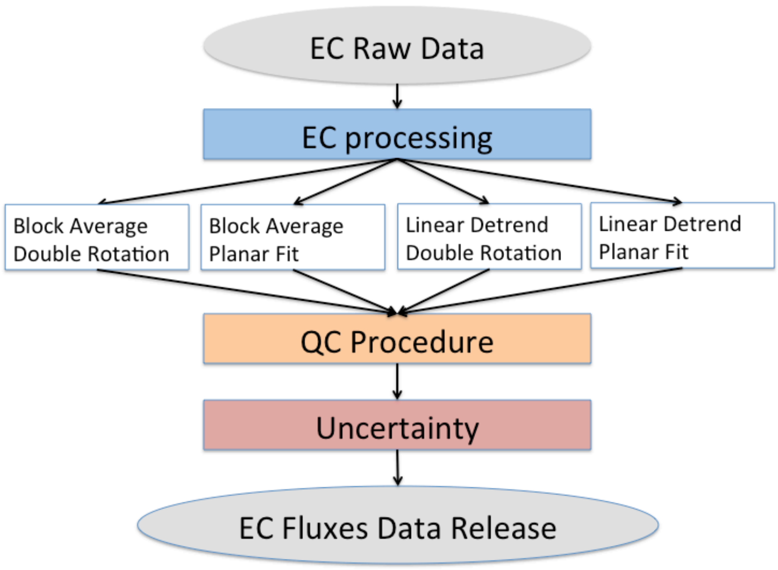

FIGURE 22. EC DATA PROCESSING PATH

EC fluxes are calculated according to four different processing schemes resulting from a combination of block average or linear detrending and double rotation or planar fit processing options. After the flux calculation step, half-hourly time series are subject to a robust and automated quality control (QC) procedure with the aim to flag and discard bad fluxes according to objective criteria. Then, uncertainty quantification is performed.

To speed up the computational time, the implementation of the four processing schemes was done in parallel mode. The whole setup, including the automated QC procedure and the uncertainty quantification, was implemented and tested in the gCube Virtual Research Environment provided by Task 7.1. By using EC raw data from a single tower, the estimated computational time required for a NRT run is of about 2 minutes.

Next Steps

The implementation of the data and metadata management is on-going and will be the focus of the next steps. In particular, the following needs to be developed and tested:

1. Data and metadata uploading tool to support RIs not strictly compliant with ICOS standardized format, including guideline on the data format
2. Data transferring from trusted repositories through the PID
3. Metadata management tool in order to automatically update the site specific input files required by EddyPro® Fortran code
4. Distributed processing for multi-site analysis in order to reduce computational time
5. Improvement of the effectiveness of the QC procedures

3.13 Use Case IC_14: SOS & SSN ontology based data acquisition & NRT technical innovations

Group Leaders

Robert Huber & Markus Stocker (UniHB)

Objectives

Research Infrastructures, specifically observatories that build on environmental sensor networks, share a common problem: data acquisition services and, in particular, the preparation of data transfer prior to data transmission are not yet sufficiently standardized. This hinders the operation of efficient, cross-RI data processing routines, e.g. for data quality checking.

The overall objective of this implementation case is to move the standardization level close to the sensors of RIs, thus allowing the implementation of common, generic data processing routines, e.g. for Near Real Time (NRT) Quality Control (QC). A further objective is to contribute to the harmonization of data transmission formats and protocols. In detail, objectives include:

- Evaluate standardized data transmission using OGC SWE Transactional SOS (Sensor Observation Service) as priority standard as well as using the Semantic Sensor Network (SSN) ontology. Both will be implemented and tested.
- Identify generic NRT QC routines suitable for multiple RIs (e.g. EMSO, EuroARGO, ANAEE, etc.).
- Identify and deploy appropriate cloud based technologies for scalable NRT QC routines.
- Implement NRT QC routines at own and/or EGI platforms and evaluate routines on raw data delivered via standardized data transmission streams.

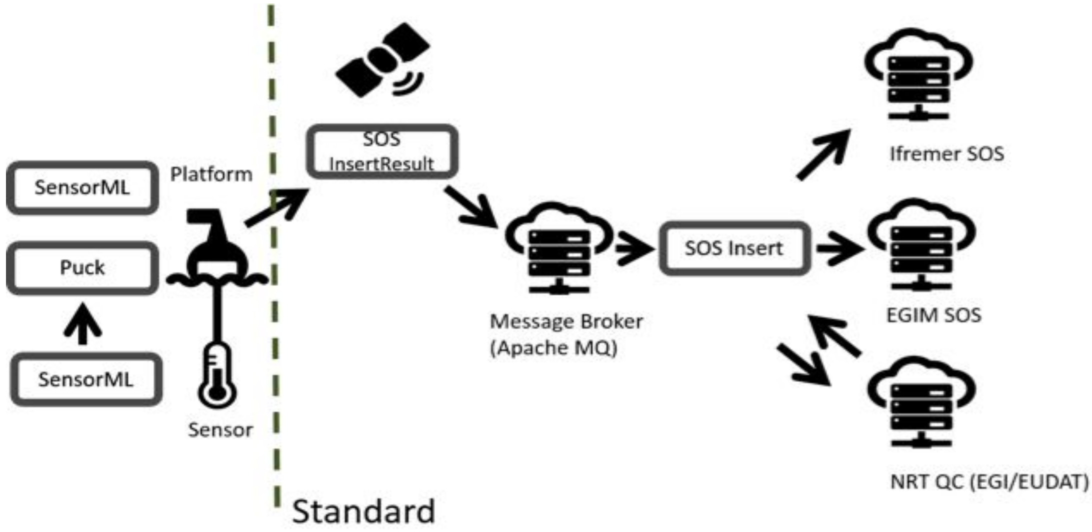

FIGURE 23. A POSSIBLE STANDARDS BASED DATA TRANSMISSION DATA FLOW. STANDARDISATION LEVEL IS BENEATH THE SENSOR AND SENSOR PLATFORM (OBSERVATORY) BASE ON OGC SWE STANDARDS. THIS IS UNLIKE MOST CURRENT ENVRIPLUS OBSERVATORIES AS THEY TRANSMIT DATA IN DOMAIN SPECIFIC OR PROPRIETARY FORMATS.

Partners

UniHB, IFREMER, CNRS, CNR, EGI, EAA, PANGAEA

Related RIs

EMSO, FixO3, EPOS, SIOS, ANAEE, SeaDataNet, EuroArgo

Activities

The proposed implementation builds on Apache Storm[41] to support scalable NRT QC on streamed standardized sensor data. Apache Storm is a distributed real-time computation system. It specializes on reliable processing of data streams and is designed to support real-time analytics and continuous computation, among other use cases. Central to Apache Storm is the notion of Storm topology. A topology consumes streams of data and processes streams in arbitrarily complex ways. It thus models the logic for a real-time application. A topology is a directed acyclic graph. Nodes are either spouts or bolts. Vertices are streams. A stream is an unbounded sequence of tuples. Tuples are data packages. A spout is a source of streams in a topology. Bolts perform computations (processing) on tuples.

We investigate the application of Apache Storm to the described implementation case. Specifically, the idea is to model the data acquisition and NRT QC computations as a Storm topology. Sensing devices may be modelled as Storm spouts, i.e. as sources of streams. Streams model the transmission of sensor data in the NRT QC application. The data – here encoded following the OGC and/or W3C standards – are modelled as tuples. Finally, any computation, e.g. for NRT QC, is modelled as a bolt of the topology.

Of particular interest are computations for NRT QC, such as outlier detection. Being modelled as bolts of a Storm topology, such outlier detection operates as a continuous computation task on the tuples of streams, as specified by the topology.

Storm topologies are deployed and executed on EGI infrastructure. For this, EGI has provided us with a cluster of virtual machines.

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

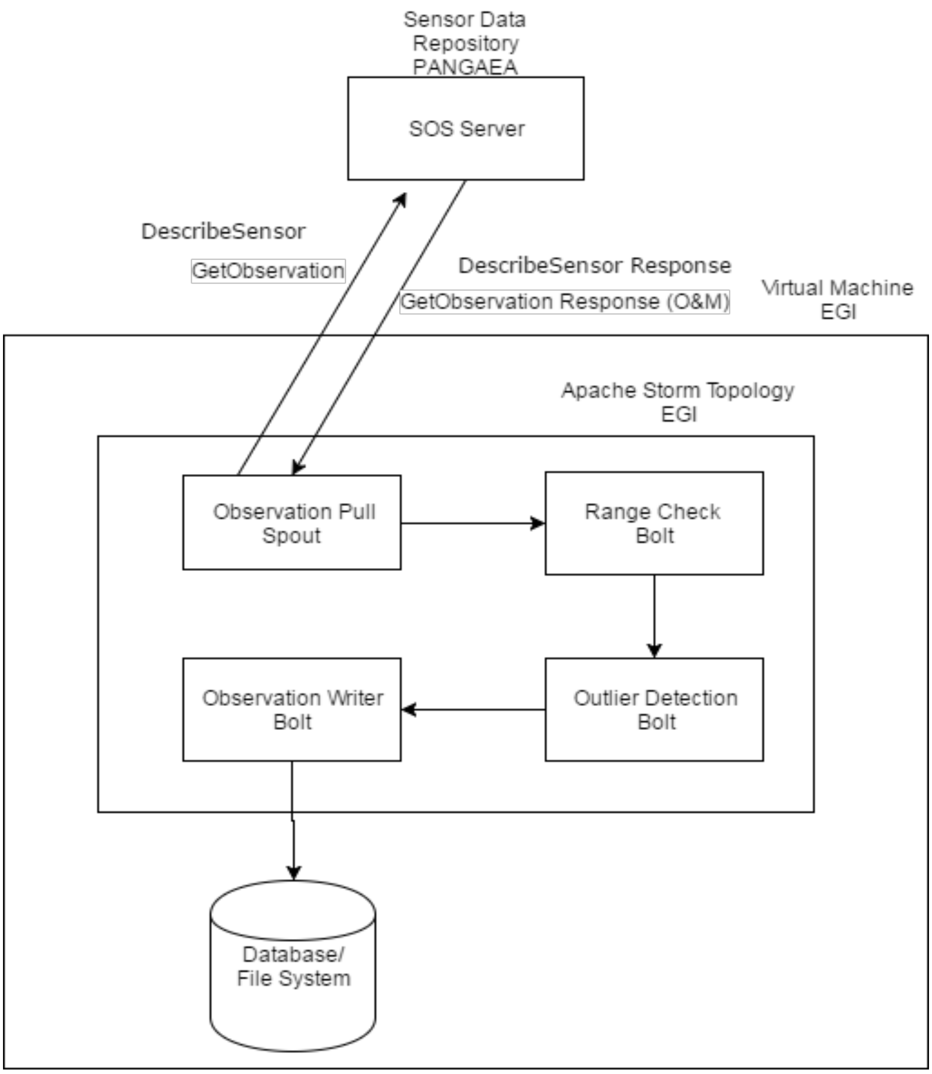

FIGURE 24. SYSTEM ARCHITECTURE SHOWING COMPONENTS AND INTERACTION. STANDARDIZED SENSOR DATA IS SERVED BY A SOS SERVER. A SPOUT REQUESTS AND EMITS OBSERVATIONS FOR QC. ONE OR MORE BOLTS IMPLEMENT QC. FINALLY, A BOLT ACTS AS A WRITER AND REQUESTS A DATABASE (OR FILE SYSTEM) TO PERSIST QUALITY ANNOTATED OBSERVATIONS.

Achievements

We have established an OGC compliant data transmission via T-SOS using the MARUM – Center for Marine Environmental Science – WaveGlider, which has been chosen as Tasks 3.3 demonstration action platform and was deployed near the Canary Islands. This data is now used for testing and has been stored in a SOS server installation at MARUM.

In collaboration with EGI, we have deployed a Storm cluster on virtual machines and tested the deployment of an example Storm topology. This enabled us to gain experience with Storm cluster deployments.

We have specified a system architecture (see Figure 24) including spouts designed to fetch observations from the SOS server and to store quality annotated observations (Observation Pull/Writer Spout) as well as bolts designed to implement QC.

Initially we use the standard O&M output delivered by a SOS GetObservation request in order to test the overall feasibility of the implementation case. Bolts will be developed that are designed for range checks and, e.g., outlier detection checks (see Figure 24).

The system for NRT QC is currently implemented in Java on EGI virtual machines with corresponding Apache Storm installation. Reporting QC results back to the SOS installation is a challenge, as the respective definition within the standard is not clear. We are currently investigating this in cooperation with experts from OGC and 52°North.

Next Steps

The implementation is on-going, in particular with respect to the development and deployment of QC routines. Furthermore, we plan to test the approach for NRT QC using the SSN ontology conformant sensor data, and compare the O&M and SSN implementations.

Having gained experiences with the implementation and deployment of NRT QC computation using Storm on an EGI cluster, we plan to discuss how the prototype may be generalized to a NRT QC service that could support other RIs. For this, we plan to consider additional data sources from other RIs.

4 SERVICE DEPLOYMENT

WP9 supports the service deployment of ENVRIplus services, and therefore, discussed with WP6-8 to understand their deployment plans and requirements. In the first part of this section, we will discuss the deployment plan and activities of ENVRIplus services development teams in WP6-8.

We have been also directly working with the use case agile teams, and prepared e-Infrastructure computing environments for service testing when this is needed. Several use cases use EGI Federated Cloud. To help other ENVRIplus service providers better understand this technology, in the second part, we will discuss service deployment approaches in this e-Infrastructure setting.

4.1 Deployment Plans of ENVRIplus Services

4.1.1 Data Identification & Citation (Services from WP6)

The objective of WP6 is to improve the efficiency of data identification and citation in the operations of Environmental and Earth Science Research Infrastructures. This can be achieved by providing access to convenient, effective and interoperable identifier management and citation services – both those developed within ENVRIplus itself or by ENVRIplus partners in other contexts. While some of these services will be very RI or domain-specific, the overarching goal is to develop tools and methods that are more broadly applicable.

Important service categories include assignment and administration of persistent identifiers for data objects throughout the research data life cycle; access to relevant metadata and retrieval of data content based on identifiers; and the collection of bibliometrics statistics for data use and citation.

In the following, we briefly describe the deployment plans and activities for a selection of WP6-related services developed by ENVRIplus partners – either in the actual framework of WP6 or in connection with other collaborations.

GBIF Integrated Publishing Toolkit and workflows for DOI assignment

ENVRIplus partner EMBRC is leading activities centered around building the GBIF[42] Integrated Publishing Toolkit – a service that seeks to assist marine biodiversity researchers with publishing their data from peer-reviewed journals, such as *Marine Biological Records*, to European data infrastructures. As an integral part, work is on-going to develop a full data lifecycle model for biological data linked to the standards developed and promoted through GBIF. The data model involves a taxonomy for known and new species, and should be able to capture geo-coded biodiversity information as well as data on physical environments and genomics.

In addition, there is a need to facilitate archiving of data from the marine biodiversity research community. To comply with international standards, and make the data FAIR[43], this will require the provision of persistent identifiers (typically DOIs) to all relevant data sets. Within the framework of the best practices defined in ENVRIplus D6.1, work is underway to define workflows for both PID assignment and data archiving.

ICOS data and metadata web services

Through its data centre, the Carbon Portal, ICOS is designing and implementing a set of services that form the backbone of the data management of the whole RI. The list includes services supporting i) data ingestion & storage, including minting of persistent identifiers; ii) staging data

---

[42] GBIF is the Global Biodiversity Information Facility, see http://www.gbif.org/ for more information.
[43] FAIR Data Principles: http://datafairport.org/

from the repository to high-performance and high-throughput computing resources; iii) easy-access cataloguing on top of an ontology-based metadata database (RDF triple store accessible via a SPARQL endpoint); iv) single-sign-on AAI (Authorization, Authentication and Infrastructure) for ICOS services; and v) a virtual research environment (VRE) platform for user-initiated data processing (based on Jupyter Notebook running on virtual machine instances). As far as possible, ICOS bases all its data management and computing services on Open Source technology.

The central component of ICOS data management is its ontology-based metadata database, which when fully developed will hold metadata about all aspects of the RI's operations – including measurement stations, people, parameter and variable descriptions, specifications of archived data objects and documents, as well as usage statistics and bibliometrics. The ICOS metadata service[44], serves many other web services, such as the ones used for data upload and of course also the data search, visualisation and download functionality of the ICOS carbon portal itself[45].

Marine RI dataset snapshot service

The ENVRIplus partner IFREMER has recently published a set of guidelines and recommendations for implementing persistent identifiers for ocean data. One important case concerns PIDs for dynamic datasets that change over time as new (observational) data become available. Central to the described approach is the maintenance of an extensive cataloguing system, that can keep track of all datasets and the dates at which data are recorded and made available. To cope with the frequent updates of datasets, the recommended approach for the involved data centres that are operating marine data repositories is to create and store separately snapshots of the database. These snapshots are however not given individual persistent identifiers. Instead, the datasets are registered (with DataCite) at their creation, and assigned a DOI that they will always keep. The snapshots are instead distinguished by appending a snapshot-specific "appendage" after the dataset DOI, preceded by a hashtag. If no snapshot identifier is appended, the landing page server will display information about the latest snapshot available.

CMIP6 PID allocation service

Presently, climate modelling projects such as CMIP6 are creating millions of files aggregated at various levels (time-series, experiments, etc.). To identify and track all data objects requires registering them at PID providers (e.g. ePIC) but the sheer number of objects means that complex technological and organizational efforts are required. In collaboration with international data federations, in particular ENES (European Network for Earth System Modelling) and ESGF (Earth System Grid Federation), Deutsches Klimarechenzentrum (DKRZ; a partner of IS-ENES) has developed a highly scalable PID registration infrastructure based on message queues.

4.1.2 Data Processing and Analysis (Services from WP7)

The data analytics platform (DataMiner) envisaged in D7.1 [Candela et al. 2017] is conceived to support the following scenarios / exploitation models:

- ***Full platform as-a-Service***: the entire DataMiner platform is operated by a service provider (e.g. D4Science) and the community / Research Infrastructure establishes a collaboration agreement to use it by well-defined service level agreements. In this scenario, the service provider can establish its own collaboration agreements with other research infrastructures to operate the service it is responsible for (e.g. this is the case of D4Science that has

[44] ICOS metadata service, described in more detail at: https://github.com/ICOS-Carbon-Portal/meta/
[45] ICOS carbon portal: https://icos-cp.eu/

established a collaboration agreement with EGI to deploy some DataMiner instances on EGI sites)

- ***Full platform as-a-Software***: a community / Research Infrastructure can decide to exploit the DataMiner technology (open source) to set up its own instance of the technology. In this case, the community / Research Infrastructure faces hardware resource costs needed to set up and operate the platform, as well as IT personnel costs needed to deploy and operate the technology. In this case, the community can set up an agreement with other infrastructures (e.g. EGI) to reduce the costs related with hardware resources. Still, the costs related to DataMiner technology deployment and operation remain
- ***Platform as-a-Service with Community Contribution***: the DataMiner core components are operated by a service provider and the platform is complemented by some instances (namely, the workers) operated by the Research Infrastructure on its own resources. This makes it possible for RIs to deploy these nodes close to where the data to be processed are actually stored or where the result data is required

Such a platform is currently operated by the D4Science infrastructure according to the "full platform as-a-Service" model. It is made available by a dedicated VRE[46] and configured to transparently exploit computing resources from EGI FedCloud (by the D4Science.org Virtual Organisation that is currently supported by CESGA, GeoGRID, IISAS-FedCloud, RECAS-BARI and UPV-GRyCAP).

The primary contribution of the optimisation topic in T7.2 (as of the time of reporting) is a micro-service based infrastructure optimisation engine, called the Dynamic Real-time Infrastructure Planner (DRIP). DRIP optimises the use of virtualised e-Infrastructure (i.e., Clouds) by automatically planning and provisioning dedicated infrastructure resources on which to run application workflows, and autonomously installing and initialising application components on those resources. DRIP is able to independently scale the infrastructure deployment based on the quality of service requirements of the application.

DRIP will be provided as a containerised service in the EGI marketplace for researchers/engineers to freely install and use for the provisioning of customised infrastructure. DRIP has already been applied on the EGI Cloud for specific use-cases, and is available to the public[47] under the Apache 2.0 open source licence. Additional micro-services to deal with specific scenarios in e-Infrastructure provisioning and customisation can be added to the DRIP service suite based on results generalised from the ENVRIplus test/implementation use cases, with the published components updated accordingly after unit and integration testing.

4.1.3 Data Curation and Cataloguing (Services from WP8)

The core of WP8 (T8.1 Curation, T8.2 Catalogue, T8.3 Provenance) is the catalog. Both curation and provenance depend upon it. The catalog is constructed by harvesting various metadata formats from the RIs (some of which are ISO19115 or its XML variant ISO19139 both of which are closely related to the EU directive INSPIRE) and converting them into a canonical form. For the purposes of ENVRIplus two canonical forms have been chosen (T8.2 D8.3): CKAN[48] as used in EUDAT and CERIF[49]. At this stage of ENVRIplus project execution, T8.2 is testing the harvesting, conversion and storage in canonical catalog form from various RIs. The deployment plan is to provide catalogs using both CKAN and CERIF to allow for respectively easy discovery and more detailed contextualisation of the resources described.

---

[46] ENVRIplus VRE in D4Science: https://services.d4science.org/group/envriplus
[47] DRIP: https://github.com/QCAPI-DRIP/
[48] CKAN: https://ckan.org/portfolio/metadata/
[49] CERIF: http://www.eurocris.org/cerif/main-features-cerif

CERIF includes role-based temporally-bounded links between entities. This means it is ideal for curation activity (T8.1 D8.1) since replicates, versions, distributed partitions / fragments with varying media storage can all be represented and linked to persons, organisations, projects, facilities, equipment, publications, etc.

For provenance, CERIF can also be used, again because of its role-based temporally-bounded links. The W3C relevant standards are PROV-DM (the provenance data model) and PROV-O (the ontology) together with various implementations such as PROV-XML (the XML schema for PROV-DM). Since the work on provenance (T8.3) is scheduled only in the latter part of the project the provenance aspects of the service are not yet defined although preliminary work has been done coordinating with T8.1 and T8.2.

For all 3 WP8 tasks the due date for the first prototype is 2018-09-30.

4.2 Approaches to Service Deployment on e-Infrastructures

In order to help ENVRIplus service providers and RIs understand how to deploy services on e-Infrastructures, WP9 maintains a guideline[50] explaining service deployment approaches. EGI is one of the e-Infrastructures supporting the ENVRIplus project and is collaborating closely with WP9. Several use cases (such as IC_14 and TC_2) and WP7 use the EGI Federated Cloud, which is a typical e-Infrastructure service allowing services to be deployed and operated. In this subsection, we use that e-Infrastructure setting as a concrete example to illustrate deployment approaches.

4.2.1 EGI Federated Cloud

EGI Federated Cloud (FedCloud)[1] is a grid of clouds with a harmonized operational behavior. EGI FedCloud federates institutional Resource Providers (RP) offering an Infrastructure-as-a-Service (IaaS) solution composed of 22 providers from 14 National Grid Initiatives (NGI) running different Cloud Middleware Frameworks (CMF), 2/3 are OpenStack, 1/3 are OpenNebula and there is one Synnefo site. Around 6000 cores are available (August 2017).

In the federation, the Clouds and their interconnections are based on open technologies such as the OpenNebula, OpenStack and Synnefo Cloud Middleware Frameworks (CMF) and on open Standards such as the Open Cloud Computing Interface (OCCI) for Virtual Machine (VM) management and the Cloud Data Management Interface (CDMI) for object storage. FedCloud is also offering native access to OpenStack sites. A common authentication and authorization layer using x509 and VOMS is used and OpenID Connect integration is on-going. Operational tools for accounting, monitoring and ticketing are operated centrally.

The EGI FedCloud value proposition is the capability of instantiating virtual machine images across heterogeneous cloud providers in a uniform way, giving the possibility to compute and store data across a combination of public and community clouds of choice and bringing computing to data.

Multiple tools are available to interact with the Federated Cloud. Firstly, the EGI Application Database is an application catalogue for Virtual Appliances (VAs) which are ready-to-run virtual machines packaged with an operating system and software application(s). Secondly, it is possible to use low-level Command Line Interfaces (CLI) and Application Programming Interfaces (API). Finally, some high-level tools such as the EGI AppDB VMops dashboard (described below) are built on those APIs to offer more user friendly and automated interfaces to FedCloud.

[50] Guideline for service deployment: https://confluence.egi.eu/display/EC/Guideline+for+Service+Deployment

4.2.2 *The EGI Application Database (EGI AppDB)*

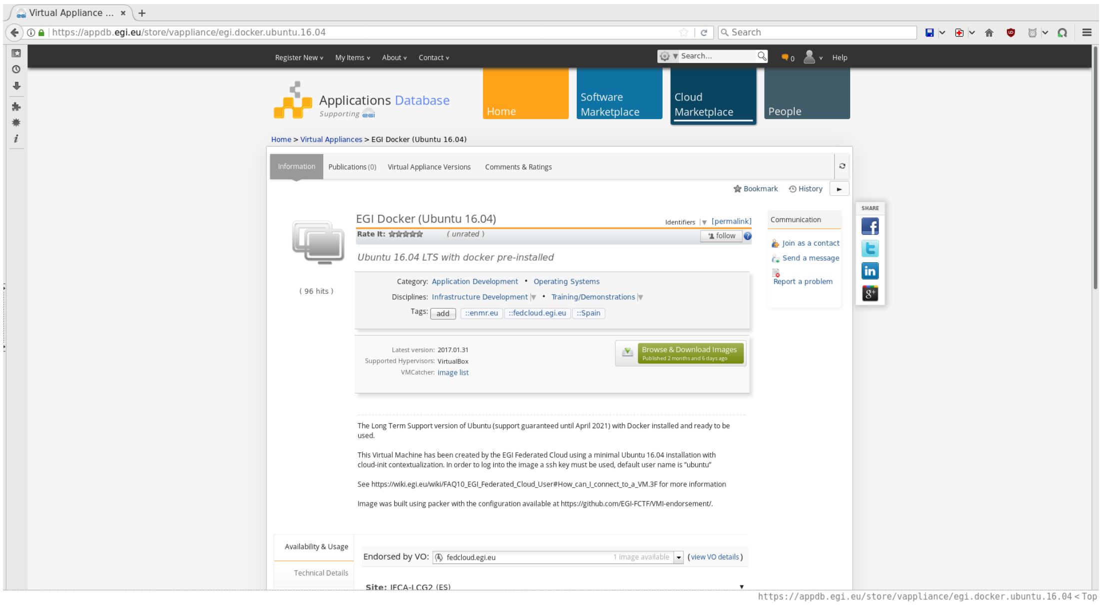

FIGURE 25: THE EGI APPDB SHOWING EGI UBUNTU DOCKER VIRTUAL APPLIANCE

The EGI AppDB[51] is an application catalogue based on a web portal for sharing:

- EGI-endorsed Virtual Appliances (VAs)
- Software packages, tarballs and scripts
- Sites and Virtual Organisation (VO) information, including VA availability at a sites for each VO
- Information about people

Using the EGI AppDB, users can share and search VAs and software. It is possible to publish metadata about the VA such as a description or hardware requirements. For the packages, software repositories are automatically created for debs and RPMs package formats, simplifying their usage. The VO-specific Virtual Appliances are synchronized on every FedCloud sites that support the VO.

4.2.3 *Deploying using Command Line Interface (CLI) and Application Programming Interface (API)*

The lowest level of interaction with EGI FedCloud is using the supported CLIs and APIs.

Currently, FedCloud is accessible using two realms: OpenStack and OCCI. OCCI can be used to access OpenStack, OpenNebula and Synnefo sites. It is also possible with OpenStack sites to use the OpenStack native tools.

Virtual Machine contextualization has to be done using cloud-init[52], it can be used to allow remote connection using ssh or to bootstrap a complete service.

4.2.4 *Deploying using Complementary Services and tools*

In addition to the IaaS cloud service, some other services are offered to the users, some of which are provided by EGI while others are just higher level services that can ease and automate the usage of an IaaS solution.

---

[51] EGI AppDB: https://appdb.egi.eu
[52] cloud-init: https://cloud-init.io/

For the long tail of science users, an Application On Demand service[53] was created, offering an easy on-demand deployment of application specific clusters (implemented using the EC3 tool) as well as some ready to use and pre-configured Science Gateways including various tools.

Some other third party applications are also able to access FedCloud. They can be grouped in two broad categories: Virtual Research Environments (VRE) and orchestrators. For example, the VREs d4science.org and WS-PGRADE are both able to use FedCloud resources and allow research developers to create user-friendly and community-based Science Gateways simplifying the use of the cloud resources.

Orchestrators are components that can automatically deploy a complete virtual infrastructure on one or multiple cloud sites, sometimes even among different cloud providers. This category includes UPV's Infrastructure Manager (IM) offering a web interface, a Command Line Interface client and multiple APIs allowing systems teams supporting a research community to easily deploy and update complex infrastructures; the Dynamic Real-time Infrastructure Planner (DRIP[54]) allowing to orchestrate deployments based on time constraints; and the INDIGO Orchestrator[55], the central component of the INDIGO Platform-as-a-Service solution. Those services can deploy infrastructure that is specified in TOSCA[56] and interact with the sites using OCCI.

4.2.5 Deploying using EGI AppDB VM Operations Dashboard (EGI VMops dashboard)

The EGI VMops dashboard[57] is a graphical interface allowing users to create and deploy topologies on FedCloud. A topology is composed of one or multiple instances of a Virtual Appliance and can include contextualization and automatic mounting of additional storage.

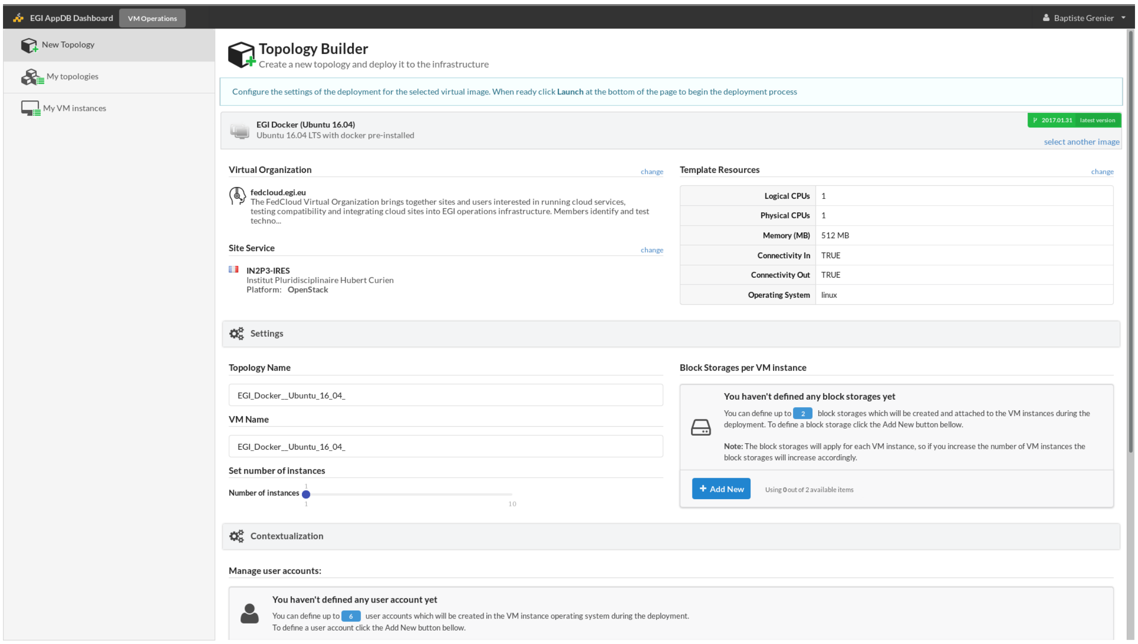

FIGURE 26: THE EGI APPDB VMOPS DASHBOARD

---

[53] Application On Demand service: https://wiki.egi.eu/wiki/AoD
[54] DRIP: https://staff.fnwi.uva.nl/z.zhao/software/drip/
[55] INDIGO Orchestrator: https://www.indigo-datacloud.eu/paas-orchestrator
[56] TOSCA: https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=tosca
[57] EGI VMops dashboard: https://dashboard.appdb.egi.eu/vmops

The EGI VMops dashboard is a web portal that is operated centrally by EGI and is built on top of the EGI services. It is integrated with the AppDB, thus allowing users to see the list of Virtual Appliances available for a specific VO and it will allow to deploy the topology on one of the available sites supporting this VO as it is also integrated with EGI Information System and monitoring. Due to the integration with the EGI CheckIn service (EGI's Authentication and Authorization Infrastructure (AAI)) there is no need for a X509 certificate for the end user. Technically the deployment orchestration is managed by the Infrastructure Manager.

The EGI VMops dashboard is an example of high-level services offering high value to users and that can be built using the standard APIs supported by FedCloud.

4.2.6 What's coming?

EGI Federated Cloud is currently focused on delivering Infrastructure-as-a-Service (IaaS) solutions, but leveraging orchestrators' work on Platform-as-a-Service (PaaS) and Software-as-a-Service (SaaS) solutions is on-going. The Service orchestration will allow the ENVRIplus community to more easily deploy and manage a complete virtual infrastructure that can be tailored to the specific needs of each Research Infrastructures. Another important effort is the OpenID Connect integration, which will simplify the access and usage of the API and CLIs by removing the need for a X509 certificate. Being used also by a number of other tools, OpenID Connect will ease integration with other solutions, such as ENVRIplus services supporting OpenID Connect, thus allowing the use of a federated Authentication and Authorization Infrastructure (AAI) supporting Single sign on across all services. Federated Cloud monitoring probes are being extended to improve the reliability of the infrastructure and exhaustively monitor it. Finally, in order to go beyond the Virtual Appliances usage and follow the growing use of containers, an EGI DockerHub of trusted containers will be created. It will offer access to selected tools useful for the various ENVRIplus RIs that use containers to easily package and deploy their user-community specific applications. They will be able to propose container configurations to easily present and share them among the EGI community.

5 SERVICE EVALUATION

5.1 Rationale

One of the major goals of ENVRIplus is to facilitate and encourage the adoption of the designs, services and data products that are being developed during the project. An important part of these activities is the active promotion of the services, including demonstrators, presentations at meetings, and the dissemination of comprehensive and user-friendly documentation, e.g. through wiki pages and manuals. However, experience has shown that to ensure uptake and, crucially, continued use beyond the lifetime of a development project like ENVRIplus, both eventual users and decision makers (in particular the thought leaders and trend setters) *must* be actively engaged throughout all phases of the service building. Indeed, without co-design and co-development the adoption is almost impossible to achieve unless it is coerced by management.

Establishing a tight and effective dialogue between everyone involved, with the explicit aim to find a common understanding of both user expectations and potential technological solutions, is a first step. To further enhance collaboration and to be able to accommodate changes in needs due to e.g. external factors or new emerging technologies, WP9 has chosen to encourage all the case development subgroups to implement agile methods, whenever possible. Although the degree and success of the adoption of agility have varied, the overall experience so far indicates that by forming productive alliances across ICT, RI and domain borders; pooling intellects and efforts; and dividing up work into concentrated drives, the case development work can be effectively steered.

Now, the challenge will be to continue along this path, and keep up the active participation of end users - as represented by RIs - in the evaluation process, starting with prototypes and first releases, and continuing throughout the lifetimes of services and other developed products.

5.2 Designing criteria and scoring methods

As outlined in the ENVRIplus Grant Agreement, all project outputs should strive to achieve a number of overarching principles, including:

- Simple but effective
- Interdisciplinary but with common rules
- Rich tools but low learning curves
- Data-driven but traceable and citable
- Scientist-centred but with a high quality of user experience
- Extensible yet robust
- Providing appropriate trust, security and privacy

Thus, in order to both help developers to appraise the usability and degree of maturity of their outputs, and to assist prospective adopters in their search for appropriate solutions, it is very important for ENVRIplus to provide standardized and easy-to-interpret assessment scores that can be used to rate and compare services according to criteria that cover as many of the listed principles as are reasonably achievable.

WP9 is responsible for all aspects of this assessment process[58], including both the design of criteria and scoring mechanism, as well as managing the evaluators and the dissemination of the results.

[58] Note that this evaluation of ENVRIplus services and products should not be confused with the assessment of cases, as described in subsection 2.3.

Evaluation criteria can be divided into those that are related to easily quantifiable parameters or outcomes, those that may be used as the basis for Key Performance Indicators (KPIs), and others which are qualitative in nature and hence more difficult to evaluate. The process to design and select useful standards and benchmarks is still on-going. In Table 5 below, we list some examples of general and ENVRIplus-specific criteria, respectively.

Following on from the detailed and localized evaluation intrinsic in the agile development processes, it is also necessary to evaluate the integrated outcomes in a timely manner. As the RIs are shaping direction and are the future adopters, it is essential that evaluation is tuned to measure their concerns. Therefore, WP9 is hosting a co-design process bringing together those in the partner RIs who are critical when adoption decisions are made, with Theme2 experts, who are developing methods, components and services.

Before any criterion can be applied, an appropriate scoring scale needs to be defined. This not only facilitates the work of the evaluators, but also ensures transparency and safeguards against bias during scoring. In this context, there remain a number of relevant issues that need to be considered in more detail, and which should be discussed further:

- How should we define scoring rules (e.g. limits or ranges, or specific alternatives)?
- Is it desirable to use
    - Individual scores for each criterion separately?
    - Combined scores (one or a few “summary numbers”)?
    - “Spider web diagrams” to help judging “by pattern”?
- Who should perform the evaluation?
    - Representatives of RIs trying to adopt or use services and products?
    - WP9 participants (RIs with PMs in Task 9.1 and/or Task 9.2)?
    - Theme2 WP and/or Task leaders, or a subgroup of these?
    - External experts?
- Should the evaluation be
    - Performed once (towards the end of the project)?
    - Repeated, for example each time a new version is released?
    - Open-ended, i.e., by providing an online scoring tool?

5.3 Plans for the evaluation process

Almost all of the evaluation criteria are relative to a certain expectation level of a desired outcome. These criteria therefore could be evaluated as a percentage of the ambition level.

The first task of the evaluation process should therefore be to quantify, for each criterion, the ambition level. A few parameters require an evaluation of the service by users. For this, a questionnaire should be developed in the next step. For sake of clarity, wherever possible evaluations should be preferably binary (satisfied/not satisfied). However, multi-level answers may be appropriate for some topics, as this allows to capture details of e.g. user experience. In addition, respondents and evaluators must be given the opportunity to provide free-text comments.

It is important to evaluate at the correct level of granularity at each stage, and to consider the needs of each category of practitioner, e.g., research scientist, research technician, research software engineer, research system administrator, data architect and data curator. Here, a challenge is to distribute evaluation forms and questionnaires in a comprehensive way, as to avoid biasing the evaluation process s by failing to adequately cover certain target groups.

The final step will be to gather the information from questionnaires and to quantify the outcomes, and to report the results per use case.

TABLE 5. EXAMPLES OF GENERAL AND ENVRIPLUS-SPECIFIC EVALUATION CRITERIA FOR SERVICES.

	Criterion	Description
General	Accessibility or Ease of use	Related to technical skill & knowledge level of adopter, quality and depth of documentation, potential need for access to specific data centres or technologies, …
	Degree of adoption outside of ENVRIplus	Number of RIs or similar organizations that have adopted a service? How many individual users?
	Maturity	Can be a scale from idea/best practice recommendation, proof-of-concept, demonstrator/pilot, operational system – compare with the Technical Readiness Level scale used by e.g. EC evaluators.
	Sustainability I	Can be assessed on many levels: Time frame of commitment by the provider, longevity of hardware/operating system/platform, longevity of software/technology, access to expertise/maintenance etc.
	Sustainability II	To what extent is the quality of engineering good enough. Is there a committed team that can grow as needed, e.g., through open-source governance, to continue development, handle maintenance and support technical questions our own (RIs') support teams cannot answer?
	Visibility	Has the service been well advertised/disseminated outside of ENVRIplus?
	Impact	Impact of the service being used or available.
	Portability	to what extent is it tied to a particular platform, toolset, language or data model. Will it embrace standards and enable innovation?
	Scope	How many aspects of the data lifecycle are involved? How many categories of expertise will this help?
	Potential	Does this direction and skeleton/conceptual framework provide a good platform on which to continue to build?
	Depth	What proportion of the technical, practical and system details does this address?
ENVRIplus-specific	Involvement of partners	How many RIs have been interested in contributing actively in the development (through agile teams or similar)?
	Domain coverage	How many domains can make use of the service?
	Scope	How many WPs are affected (e.g. how many "pillars" and "cross beams" are related)?
	Explicit ENVRIplus contribution	Would the service have been developed anyway, or was ENVRIplus instrumental in the process?
	Adoption by partner RIs	Number of ENVRIplus partners that have adopted the service (in whole or in part). How many individual users?

6 CONCLUSIONS AND FUTURE WORK

This report summarizes the progress on the service integration and deployment issues explored in Task 9.1. It describes the method and process of collecting community use cases, and the agile approach for design, implementation and maintenance. The agile teams play an important role in testing and validating of ENVRIplus service solutions. Most of their members are researchers directly involved in RI communities, and have good knowledge of community requirements. Through agile activities, they help to establish collaborations across RIs, WPs, and organizations. The agile teams work very efficiently, and several made important progress within short time. A showcase session was organized by WP9 during the 4th ENVRIweek, where 4 use cases (TC-2, IC-3, IC_13, IC_14) gave demonstrations to the whole ENVRI community and shared their development results and experiences. Among them, use case TC_2 was selected as one of three Theme2 highlights and presented in the project review meetings.

In future work, T9.1 will continue coordinating the agile activities towards final demonstrators. Agile teams will be encouraged to consider how to generalise their approaches to support more ENVRIplus RIs, and test their service solutions with new RI communities. However, an agile team knows its own user and technical context better than anyone, and provided it considers goals and priorities set by the leadership and management, it may choose the focus of its next big push. We will follow the evaluation plan to evaluate the outputs of the integration services.

In order to better manage ENVRIplus service development and delivery, both from Theme2 WPs and use cases, T9.1 is also working on ENVRIplus Service Portfolios[59], including a list of services that ENVRIplus offers to potential requesting customers. A Service Portfolio provides information about the planning, design, delivery, and on-going assessment and refinement of services through their entire lifecycle. It gives a better understanding of the components, dependencies and processes behind service delivery; helps assessing what works and what doesn't; and gives greater insight into the impacts that ENVRIplus services have.

We will work together with T9.2 to promote the ENVRIplus solutions, and in particular explore how to connect ENVRIplus RIs with the European Open Science Cloud.

As part of the European Commission Digital Single Market strategy[60], the European Open Science Cloud (EOSC) initiative was officially launched in April 2016 by the European Commission. EOSC promotes scientific excellence and data reuse, and drives Europe-wide cost efficiencies in scientific infrastructures through the promotion of interoperability on an unprecedented scale.

EOSCpilot was funded as the first project in this area, and the main focus is on **a governance and business model**. At the time of finishing this deliverable[61], we have received good news that the EOSC-Hub proposal has been evaluated favourably. Led by EGI, the EOSC-Hub project will build an **e-Infrastructure foundation** for the European Open Science Cloud. It aims to integrate technology and solutions from 3 European e-Infrastructures (EGI, EUDAT and INDIGO) to provide an interoperable computational platform that supports easy access to data and computing resources for European scientific researchers. Following that, the EC will announced the thematic clouds calls to **bring European domain scientific research RIs onto the Open Science Cloud**. This includes INFRAEOSC-04-2018, Connecting ESFRI infrastructure through Cluster projects, which targets communities such as ENVRI.

[59] ENVIRplus Service portfolios: https://confluence.egi.eu/display/EC/ENVRIplus+Service+Portfolios
[60] European Commission (2015), Open Science at the Competitiveness Council. http://ec.europa.eu/digital-agenda/en/news/open-science-competitiveness-council-28-29-may-2015
[61] In Jul 2017

If the ENVRI community is going to respond to this call, WP9 will be in a good position to contribute to the proposal preparation. Two of Europe's leading e-Infrastructures, EGI and EUDAT, participate in WP9, and several use cases have already explored the interfaces between ENVRIplus RIs and e-Infrastructures. On the other hand, WP9 has initiated a discussion in the workshop "Towards an *ENVRI-as-a-Service to European Open Science Cloud*"[62] in DI4R conference 2016, Krakow. The conference[63] was organized jointly by major European e-Infrastructures, EGI, EUDAT, GÉANT, OpenAIRE, PRACE and RDA Europe.

In the next step, use of e-Infrastructure services and resources will be further explored via WP9 agile teams, for example in the areas of:

- Compute-intensive processing
- Data-intensive storage and processing
- Integration of e-Infrastructures' services to provide new functionality
- Publishing and sharing data and services using ready-to-use e-Infrastructure services
- Shipping data and services across countries – specially useful for those RIs that do not have the capacity
- Enabling Open Science by federating data and services for all disciplines
- Replication and backup of data and services
- Hosting of web services
- Improving accessibility by providing nearer access sites
- Combinations of all above

These activities and experiences will facilitate the positioning of ENVRIplus RIs within the EOSC.

REFERENCES

[Atkinson 2015] M. Atkinson, M. Carpené, E. Casarotti, S. Claus, R. Filgueira, A. Frank, M. Galea, T. Garthk, A. Gemünd, H. Igel, I. Klampanos, A. Krause, L. Krischer, S. Hoon Leong, F. Magnoni, J. Matser, A. Michelini, A. Rietbrockk, H. Schwichtenberg, A. Spinuso and Jean-Pierre Vilotte: VERCE delivers a productive e-Science environment for seismology research. In Proc. IEEE eScience 2015. Doi 10.1109/eScience.2015.38

[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

[Candela et al. 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, March 2017

[Rauber 2015] A. Rauber et al., "Data citation of evolving data. Recommendations of the Working Group on Data Citation (WGDC)". Preliminary report from 20 Oct 2015. Available at https://rd-alliance.org/system/files/documents/RDA-DC-Recommendations_151020.pdf Accessed 2017-01-30.

[Rauber 2016] A. Rauber, A. Asmi, D. van Uytvanck and S. Pröll, "Identification of Reproducible Subsets for Data Citation, Sharing and Re-Use". *Bulletin of IEEE Technical Committee on Digital Libraries*, vol. 12, issue 1, May 2016, 6-15. Available at http://students.cs.tamu.edu/ldmm/tcdl/v12n1/papers/IEEE-TCDL-DC-2016_paper_1.pdf

[62] Towards an *ENVRI-as-a-Service to European Open Science Cloud*" in DI4R conference 2016, Krakow https://www.digitalinfrastructures.eu/content/towards-environmental-research-infrastructure-service-open-science-cloud

[63] DI4R conference: https://www.digitalinfrastructures.eu/

APPENDIX A. USE CASE DESCRIPTION TEMPLATE

<Title of the use case>

1. Background

1.1 Short description

<Short description of the use case.>

1.2 Contact

Background	Contact Person	Organization	Contact email
<Choose one of the following roles: [RI-ICT | RI-Domain | ICT | e-Infrastructure]>	<Full name>	<Organization of the contact person>	<Email>

1.3 Use case type

<Please indicate the use case type based on the following description:

Scientific use case

Conditions:

- A real research project dependent on various new RI services (methods);
- Frontier international research with excellent scientists;
- The science case serves as a story to sell ENVRIPLUS.

Implications:

- The project leader should be willing to serve as a guinea pig and test solutions;
- The use case involves one of the research institutions participating in ENVRIPLUS;
- Funding should be provided for a postdoc in the science case.

Test case

Conditions:

- The test case is build an on a new and developing RI service, so that ENVRIPLUS staff can understand its rationale;
- The test case covers topics of relevance for various WPs, such as instrumentation, data flows, and training;
- The test case is part of the RI's portfolio of implementation cases.

Implications:

- The involved RIs have enough developers to work together with ENVRIPLUS staff;
- The test developments to be shared with ENVRIPLUS partners.

Implementation case

Conditions:

- Each RI describes its portfolio of new and/or enhanced services that they expect from ENVRIPLUS results, derived from the ENVRIPLUS WPs;
- ENVRIPLUS staff work with the RIs on these descriptions, which in the course of the project will be gradually updated with more details.

Implications:

- Implementation cases selected and adopted by interested RIs;
- Both RIs and ENVRIPLUS invest in the actual implementation and associated services.>

1.4 Scientific domain and communities

Scientific domain

<Please indicate one of more sub-domains in the environmental and earth sciences [Atmosphere | Biosphere | hydrosphere | geosphere].>

Community

<Please choose one or more close communities of this use case: [Data Acquisition | Data Curation | Data Publication | Data Service Provision | Data Usage | Or others]. Please check http://confluence.envri.eu:8090/display/ERM/SV+Communities for detailed explanation of those communities.>

Behaviour

<please identify the relevant community behaviour of this use case based on the model in http://confluence.envri.eu:8090/display/ERM/SV+Community+Behaviours. Please highlight any important roles e.g., as suggested in http://confluence.envri.eu:8090/display/ERM/SV+Community+Roles >

2. Detailed description

Objective and Impact

<Please describe the main objectives of the use case and its expected impact technically and scientifically.>

Challenges

<Please describe any particular challenges and difficulties of the use case that you think need to be faced.>

Detailed scenarios

<Please describe the basic scenarios supported by this use case—diagrams which can indicate flows among processes or activities are particularly welcome.>

Technical status and requirements

<Please describe the possible components in what detail you can, considering both functionality and current status (e.g. does it involve existing tools/data?). Please also consider the requirements for the use case.>

Implementation plan and timetable

<Please propose an agenda for implementing the case. This need not be particularly detailed at this point in time, but some aspects worth considering include: 1) timeline, 2) milestones, 3) involved RIs, 4) links to ENVRIplus work packages / tasks, 5) allocation of resources (staff time and budgets) from RIs and ENVRIplus. >

Expected output and evaluation of output

<Please describe the expected output for implementing the case, and the measurable criteria by which the output will be deemed a success.>

APPENDIX B. USE CASE REVIEW FEEDBACK FORM

Instructions:

- The purpose of use case review is to 1) help case proposer to improve the 'shape' of the use cases, 2) identify the development feasibility, 3) suggest the development priorities.
- A good-shape-use-case needs1) to be feasible for implementation and deployment, 2) well demonstrate the value of ENVRI+ e.g., supporting multidisciplinary research, and help sell ENVRIplus services, 3) have high scientific value, and 4) express wide community interests.
- The cases will be reviewed from four viewpoints: 1) scientific significance, 2) function demonstration, 3) implementation feasibility, and 4) deployment feasibility. For different types of the use cases,
    - Evaluation areas [3, 4] are mandatory for Implementation Cases, others are optional
    - Evaluation areas [2,3,4] are mandatory for Test Cases, others are optional
    - Evaluation areas [1,2] are mandatory for Science Cases, others are optional
- Review deadlines: according to the agreement in https://docs.google.com/spreadsheets/d/1KtMIC3UfaVxXZINbbXGR_TuNTvN-Ogh9Tsfdhhluj2Y/edit#gid=0

For scoring { 1 ⇒ very weak, 2 ⇒ weak, 3 ⇒ good, 4 ⇒ very good, 5 ⇒ excellent }

[Evaluation Area 1]: *Does this use case have high science significance (in single or multiple domains)? If not what modifications do you suggest?*

- Reviewer:
- Your score [1-5]:
- Your comments:

[Evaluation Area 2]: Does this use case demonstrate ENVRIplus value well? If not what modifications do you suggest? Things to be considered: specific services developed in theme2, interoperability, reusability, etc.

- Reviewer:
- Your score [1-5]
- Your comments:

[Evaluation Area 3]: Is this use case feasible for implementation? If not what modifications do you suggest? Please also provide answers to the following questions:

- ***Does the use case description include a timetable with an end date?***
- ***Does the use case clearly estimate the development effort and required resources?***
- ***Can the use case description achieve a valid target in the indicated time period?***
- ***Should parts be omitted for time reasons or to reduce complexity?***
- ***Do you have suggestions for any additional expertise (names)?***

1) Reviewer:
2) Your score [1-5]
3) Your comments:

[Evaluation Area 4]: Is this use case feasible for deployment? If not what modifications do you suggest?

- Reviewer:
- Your score [1-5]:
- Your comments: