D9.2
Service deployment in computing and data e-Infrastructures Version 2

WORK PACKAGE 9 – Service Validation and Deployment

LEADING BENEFICIARY: EGI Foundation

Author(s):	Beneficiary/Institution
Yin Chen	EGI Foundation
Ingemar Haggstrom	EISCAT Science Association
Justin Buck	BODC
Markus Stocker	TIB
Thierry Carval	EuroArgo/Ifremer
Domenico Vitale	University of Tuscia
Robert Huber	University of Bremen
Margareta Hellstrom	ICOS/Lund
Leonardo Candela	CNR-ISTI
Florian Haslinger	ETHZ

Accepted by: Yannick Légre (WP 9 leader)

Deliverable type: [DEMONSTRATOR]

Dissemination level: PUBLIC

Deliverable due date: 31.8.2018/M40

Actual Date of Submission: 31.8.2019/M40

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 Sep 2017 (M29) to Aug 2018 (M40). It is an update of a previous deliverable D9.1. It reports the implementation results of community use cases for testing and validating ENVRIplus Theme 2 service solutions.

Project internal reviewer(s):

Project internal reviewer(s):	Beneficiary/Institution
Paul Martin	University of Amsterdam
Vito Vitale	CNR-ISAC
Anca Hienola	FMI

Document history:

Date	Version
1.6.2018	Draft for comments
30.7.2018	Completed version for internal review
31.8.2018	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

This document reports on implementation results of T9.1 use cases for testing and validating ENVRIplus Theme 2 service solutions. It is an update of the previous report D9.1 that was submitted on M28. In D9.1 we explained the approach to the service integration and validation. The driver was to define appropriate use cases that address community's real needs, integrate ENVRIplus Theme2 services as part of the service solutions, and implement those use cases following an agile approach.

This document describes 7 well-developed use cases that are selected as the final Science Demonstrators. By Science Demonstrator we mean "a showcase of a service solution illustrated through a prototype implementation, which serves as proof or evidence that the Theme2 services can bring added value for supporting ENVRIplus community to deliver scientific research".

Science Demonstrator 1 addresses a requirement of the EISCAT RI community, namely to allow individual scientists to process their experimental data using their own algorithms. The challenge is common to many ENVRIplus RIs, where data is often processed using standard models and methods. As researchers want to use different analysis models, easily modify parameters or algorithms, and collaborate with each other, they need a Virtual Research Environment (VRE). This demo showcases a model making use of the D4Science gCube platform developed by T7.1, which enables scientific researchers to re-process data by implementing and adapting algorithms and parameters from other sources.

Science Demonstrator 2 showcases a novel implementation of a computationally efficient tool for processing of Eddy Covariance (EC) data which offers to users the possibility to calculate EC fluxes through the EddyPro® software (LI-COR Biosciences, 2017; Fratini and Mauder, 2014) according to 4 processing schemes resulting from a different combination of existing methods. To reduce the computational runtime required, the 4 processing schemes were implemented and executed in parallel mode. The whole service setup including a metadata management algorithm, was implemented and tested in the D4Science gCube Virtual Research Environment provided by Task 7.1, and the final computational runtime for Near Real Time (NRT) processing (i.e. flux estimates based on raw data collected the previous day) is of about 4 minutes, similar to those required for a standard run involving only a single processing scheme.

Science Demonstrator 3 addresses a common problem for ENVRIplus RIs (specifically observatories that build on environmental sensor networks) that data acquisition service, in particular, the preparation of data transfer prior to data transmission are often not yet sufficiently standardized. This hinders the operation of efficient cross-RI data processing routines, e.g., for data quality checking. The demonstrator showcases a service prototype that allows submitting and publishing raw observational (non-geophysical) environmental time series data in common standard formats (T-SOS XML and SSNO JSON). A messaging API (EGI ARGO) is used to perform Near Real Time (NRT) quality control procedures by an Apache Storm NRT QC Topology, which publishes the quality controlled and labelled data via a messaging output queue.

Science Demonstrator 4 describes the EuroArgo Data Subscription Service (DSS) that allows researchers to subscribe to customized views of Argo data, selecting specific regions and time-

spans, and choose the frequency of updates. Tailored updates are then provided on schedule to researchers' private storage. The demo showcases an integration solution that combines the EuroArgo community data portal with e-Infrastructure services (EUDAT B2SAFE, EGI FedCloud, etc.), and uses the DRIP service developed by T7.2 for optimised service deployment. The pilot activity was initiated by the marine research community, however, the possibility to receive regular transmissions of data, especially in near-real time, directly from the organisation responsible for data collection and (pre-)processing, is very important to many large initiatives. ENVRIplus RIs can benefit from the subscription services, e.g., to create more elaborated data products by requesting data from other sources, and can optimise their internal workflows by signing up for automatic updates.

Science Demonstrator 5 showcases a "sensor registry" that aims at supporting the management of sensors deployed for in-situ measurements. Common sensors or families of sensors are used across different research infrastructures, for example, oxygen optodes that are equipped on platforms in multiple research infrastructures. The goal of this work is to define common methods to access the sensor metadata in such cases. The sensor registry applies the design principle of data catalogue developed in WP8, and uses data technologies and standards from the OGC Sensor Web Enablement family including SensorML, Observations and Measurements (O&M), and Sensor Observation Service (SOS). It brings together a marine domain implementation of these standards (the Marine SWE profile) developed by several European projects demonstrating the viability for future sensor and observation activities. The service can be integrated to various types of platforms, deep-sea observatories (e.g., EMSO), marine gliders (e.g., EuroGOOS) as well as solid earth (e.g., EPOS) or atmosphere observations (e.g., ICOS). It can also be used to track usage of specific sensor models (e.g., CO2) across the RI 's observation networks.

Science Demonstrator 6 describes a service prototype that supports aerosol scientists in studying new atmospheric particle formation events by moving data analysis from local computing environments to interoperable infrastructures, thus harmonizing data analysis itself and more importantly the syntax and semantics of data derived from analysis. As researchers interpret primary data and thus gain information and transfer information into knowledge, we are studying and advancing in particular some technical aspects of a knowledge infrastructure i.e., a robust network of scientists, artefacts such as virtual research environments and research data, and institutions such as research infrastructures and e-Infrastructures that acquire, maintain and share scientific knowledge about the natural world. The science demonstrator showcases a possible architecture of a socio-technical infrastructure that "transforms data into knowledge." The proposed approach highlights a range of novel possibilities, in particular enabling researchers to focus on data analysis and interpretation while leaving data access and transformation from and to systems to interoperable infrastructure. It significantly contributes to implementing the global agenda of FAIR data by promoting the notion of "FAIR by Design", weaving data FAIRness into the fabric of infrastructures. It builds on the principle not to leave making data FAIR to researchers but to guarantee it by design of well-engineered infrastructures. The demonstrator is first and foremost of primary interest to a specific scientific community, namely the various aerosol research groups that study new particle formation events.

Science Demonstrator 7 illustrates how a LifeWatch researcher can easily upload and integrate an analysis algorithm in D4Science, and share it with other researchers in a VRE. The use case

proposed an integration solution that links the D4Science/gCube VRE to the LifeWatch RI and to the EGI e-Infrastructure. This integration, for example, enables individual researchers to repeat and reuse algorithms at will, run trend analysis, and add new parameters and custom data. The VRE provides provenance registration that improves reproducibility and also allows retention of computation results in the user's workspace. This facilitates editing and adaptation of algorithms, features that are not provided by the existing LifeWatch ICT.

1 Introduction

1.1 Scope and Purpose

The aim of ENVRIplus WP9 is to validate the results of the ENVRIplus Theme2 developments by showcasing software deployments onto computing and data infrastructure and through investigating how to operate developed services within RIs. WP9 defined two tasks:

- T9.1 focuses on the technical issues of software validation, integration and release management, and of deploying developed results on computing and data infrastructure
- T9.2 focuses on tracking the actual usability and operationality of the ENVRIplus services, once these are deployed under real world conditions.

This deliverable reports Tasks T9.1 activities during the project period M29 to M40. It is an update of the previously submitted deliverable D9.1, and describes the implementation results of use cases identified in that report.

1.2 Approach

In D9.1[1] we explained the approach to the service integration and validation. The driver was to define appropriate use cases that address community's real needs, integrate ENVRIplus Theme 2 services as part of the service solutions, and implement those use cases following an agile approach.

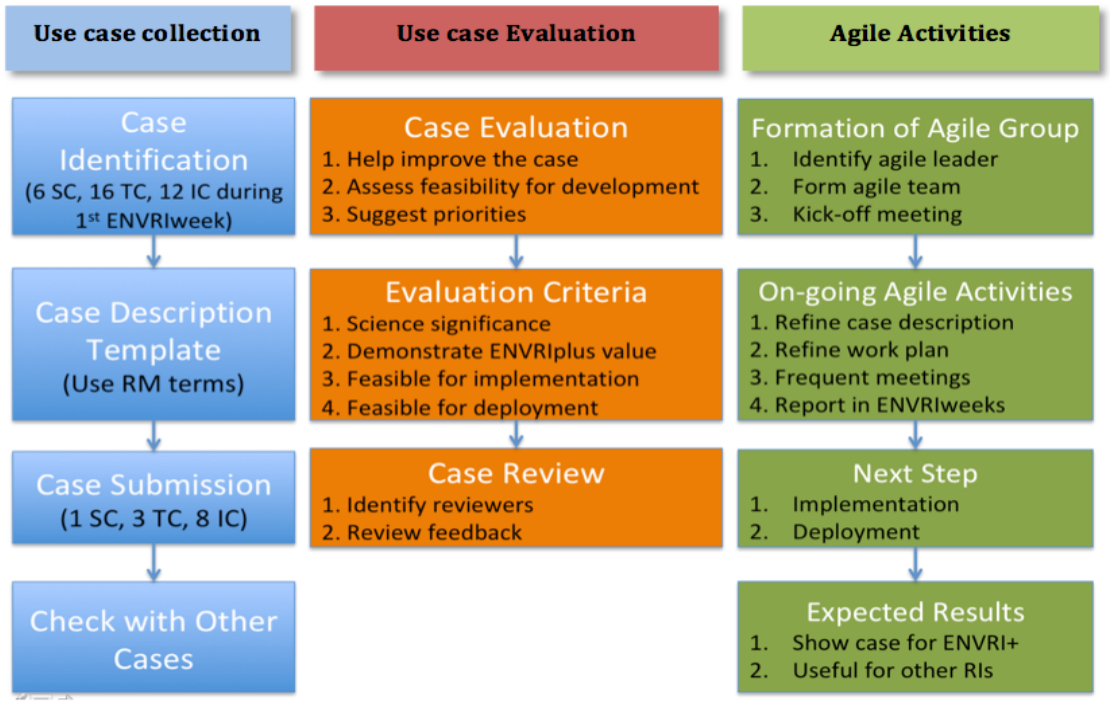


FIGURE 1. THE THREE MAIN STEPS OF THE USE CASE COLLECTION PROCESS: 1) USE CASE COLLECTION, 2 USE CASE EVALUATION AND 3) AGILE ACTIVITIES

The process from use case identification to implementation is summarized in Figure 1, and was described in more detail in D9.1. In Step 1) *Use Case Collection*, we opened a call to all Theme 2

[1] ENVRIplus deliverable D9.1: http://www.envriplus.eu/wp-content/uploads/2015/08/D9.1-Service-deployment-in-computing-and-internal-e-Infrastructures.pdf

partners for use cases submission. We requested that the use cases should be identified together with ENVRIplus RIs, and the case description should include the investigation goals, proposed solutions for integrating Theme 2 services, and a feasible work plan to achieve the goals. In Step 2) *Use Case Evaluation*, each submitted use case was evaluated by two reviewers on aspects of scientific significance, demonstration of ENVRIplus value, and feasibility for implementation and deployment. The finally selected use cases are listed in Table 1. Finally, in Step 3) *Agile Activities*, each selected use case was implemented following the agile methodology. An implementation team ('*agile team*') was formed and team leaders were identified. The agile teams started with refining the case descriptions, then worked according to defined plans, and reported back at ENVRI weeks. By M28, when D9.1 was written, the formation of agile teams was completed. Since M29, T9.1 activities are focusing on use case implementations.

TABLE 1. SELECTED USE CASES, WITH HIGHLIGHTED THEME 2 FINAL SCIENCE DEMONSTRATORS

No.	USE CASES	INTEGRATED SERVICES[2]	AGILE GROUP LEADERS
IC_1	Dynamic data citation, identification & citation	Identification & citation (WP6)	Alex Vermeulen & Margareta Hellstrom (LU)
IC_2	Provenance Implementation case	Provenance (WP8)	Barbara Magagna (EAA)
IC_3	**User support to re-process data using their own algorithms**	**gCube Data Analytics facility (WP7), D4Science VRE**	**Ingemar Haggstrom (EISCAT), Leonardo Candela (ISTI-CNR)**
IC_9	Quantitative accounting of Open Data use	Identification & Citation(WP6)	Markus Fiebig (NILU)
IC_10	Domain extension of existing thesauri	Semantic Linking (WP5)	Barbara Magagna (EAA)
IC_11	Semantic Linking Framework	ENVRI RM (WP5), Catalogue (WP8), Semantic Linking(WP5)	Zhiming Zhao & Paul Martin (UvA) Barbara Magagna (EAA)
IC_12	Implementation of ENVRI(plus) RM for EUFAR and LTER	ENVRI RM(WP5)	Barbara Magagna (EAA)
IC_13	**The eddy covariance fluxes of GHGs**	**gCube Data Analytics facility (WP7), D4Science VRE**	**Dario Papale & Domenico Vitale (UNITUS)**
IC_14	**SOS & SSN ontology based Data Acquisition and NRT Data Quality checking services**	**Semantic Linking (WP5), Sensor Observation Service, EGI ARGO messenger (WP9), EGI FedCloud (WP9), EGI storm(WP9)**	**Robert Huber & Markus Stocker (UniHB)**
TC_2	**EuroArgo Data subscription service**	**DRIP (WP7), EUDAT B2FIND (WP9),EUDAT Data Subscription Service (WP9), EGI FedCloud (WP9)**	**Thierry Carval (IFREMER), Yin Chen (EGI)**

[2] For services directly stemming from Theme 2 work packages is reported in parenthesis.

TC_4	**Sensor registry**	**Data Catalogue, Provenance (WP8)**	**Justin Buck (BODC)**
TC_16	Description of a National Marine Biodiversity Data Archive Centre	ENVRI RM (WP5)	Dan Lear (MBA), Abraham Nieva & Alex Hardisty (CU)
TC_17	**New particle formation event analysis on interoperable infrastructure**	**gCube Data Analytics facility (WP7), D4Science VRE, EGI JupyterLab (WP9), gCube/D4Science Catalogue(WP8), ENVRI RM (WP5)**	**Markus Stocker (TIB)**
SC_3	**gCube-based VRE for Mosquito disease study**	**gCube Data Analytics facility (WP7), Provenance (WP8)**	**Matthias Obst (UGOT), Baptiste Grenier (EGI)**

1.3 Integration and Validation

Use cases are implemented following an agile development approach. Agile development refers to a set of software development methods in which requirements and solutions evolve through collaboration between self-organising, cross-functional teams. Establishing short-lived, agile multidisciplinary task forces to address specific issues is key to developing the depth of understanding and commitment to make progress[3]. The agile teams are responsible for testing technologies, service integration and deployment. Most of the agile team members are researchers directly involved in ENVRIplus RIs, with good knowledge of respective community requirements. They bring information and feedback from their RIs and at the same time introduce ENVRIplus services and development results to their communities.

Table 1 also provides the mapping of the use cases and tested Theme 2 services. The detailed integration solutions are discussed in each Science Demonstrator in Section 2. During ENVRIweeks we organised sessions or workshops where agile teams discussed their use cases and got feedback from the communities.

- 1st ENVRIweek, *16-20 Nov 2015, Prague*: A plenary session was organized to involve all ENVRIplus RIs and WPs to jointly identify use cases that would be of interest for community researchers and benefit their daily work. This also opened up conversations among Themes and WPs to establish collaborations and to avoid duplication of effort.
- 2nd ENVRIweek, *9-13 May 2016, Zandvoort*: A use case session was organized involving both Theme 1 and Theme 2 members. Representatives of 14 reviewed and selected use cases gave presentations on their investigation plans. The purpose was to form agile teams involving both domain scientists from Theme 1 and technology experts from Theme 2.
- 3rd ENVRIweek*, 14-18 Nov 2016, Prague*: To contribute to the Theme 2 main discussions at that time on architecture design and integration of e-Infrastructure services, we organised a session during the Theme 2 mini workshop that was attended by Theme 2 and e-Infrastructure representatives. Use cases that tested e-Infrastructure services

[3] M. Atkinson, M. Carpené, E. Casarotti, *et al.*: VERCE delivers a productive e-Science environment for seismology research. In Proc. IEEE eScience 2015. Doi 10.1109/eScience.2015.38

were presented (including TC_2 EuroArgo data subscription service, TC_13 ICOS data processing use case, IC_3 EISCAT-3D use case on supporting individual researchers for data processing, and SC_2 LifeWatch use case on mosquito study). The discussions focused on using e-Infrastructure technology and architecture issues. Theme 2 developers and e-Infrastructure service providers gave feedback on service integration and e-Infrastructure solutions.

- 4th ENVRIweek, *15-19 May 2017, Grenoble*: A Theme 2 showcase workshop was organized, targeting the whole ENVRIplus community. Four well-developed use cases (IC_3, IC_13, IC_14, and TC_2) were demonstrated and feedback from RIs communities was received. Following that, a discussion session was organised on how to make the approach more generic so that it would also support other ENVRIplus RIs, taking the four use cases as examples. We also identified RIs that could be further involved to test and validate the service solutions.
- 5th ENVRIweek, *6-10 Nov 2017, Malaga*: In line with the European Open Science Cloud (EOSC) discussions at that time, the session agenda included presentations from selected use cases (IC_3, IC_9, IC_13, IC_14, TC_4 and TC_17) that addressed EOSC issues (e.g., open access, FAIRness, interoperability, user-driven, using e-Infrastructure resources and services, etc.) and discussions on how to connect ENVRIplus services with EOSC. Positive feedback was received from EOSC (EOSCpilot and EOSC-hub) project representatives that showed these pilot investigations were among the pioneers implementing the EOSC vision and that they could serve as examples to help ENVRIplus RIs connect to EOSC. On the other hand, the use cases also identified the requirements from environmental science community for EOSC, thus benefitting EOSC developments.
- 6th ENVRIweek, *14-18 May 2018, Zandvoort*: Due to the tight schedule , instead of holding a dedicated WP9 session,, we organised a series of WP9 webinars afterwards to catch up on the missed discussions. Each online webinar focused on one use case, which gave sufficient time for detailed discussions. In particular, the use cases served as vehicles to reach RIs by asking the use case leaders to invite Theme 1 colleagues and end users. These webinars attracted good attention, and the use cases benefitted from various feedback and inputs.

In order to provide quantitative measures of validation information and make it available to service developers, service providers and end users, T9.1 supports T9.2 to propose and develop an new approach that defines a soundness evaluation criteria matrix, and developing a service evaluation tool with special focus on user friendliness, aspects covered and possibilities for customising both the user interface and the output formats. The detailed description will be provided in the upcoming WP9 deliverable D9.4.

1.4 Deployment

WP9 provides e-infrastructure environments for service deployments. Two leading European e-Infrastructures participate in WP9, namely EGI[4] and EUDAT[5]. 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.

[4] EGI: www.egi.eu
[5] EUDAT: www.eudat.eu

EGI Federated Cloud (FedCloud) resources are offered for testing and deploying Theme 2 services and the pilot implementations from use cases. EGI (FedCloud)[6] is a grid of clouds with a harmonized operational behavior. EGI FedCloud federates institutional resource providers 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 OpenStack, 1/3 OpenNebula, and 1 Synnefo site. Around 7000 cores are available (as in August 2018).

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. 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 data analytics platform (DataMiner) envisaged in T7.1 is operated by the D4Science infrastructure according to the "full platform as-a-Service" model. It is made available by VREs[7] 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. A Service Level Agreement (SLA) has been signed between EGI and these resource providers that guarantees the provision of the resources and the quality of the service.

The Dynamic Real-time Infrastructure Planner (DRIP) developed in T7.2 optimises the use of virtualized e-Infrastructures (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 has been deployed and tested on the EGI FedCloud, and is available to the public[8] under the Apache 2.0 open source licence. Additional micro-services dealing 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.

Several use cases integrate e-Infrastructure services as part of the service solutions, and deploy the final systems onto the e-Infrastructure resources provided by T9.1. For example, in use case IC_14, the Near Real Time (NRT) Quality Control (QC) service is implemented in Java on EGI virtual machines with a corresponding Apache Storm installation. In use case TC_2, accumulating monthly EuroArgo datasets are pushed to EUDAT e-Infrastructure (B2SAFE[5]), and synchronized between EUDAT and EGI FedCloud. The process was executed by DRIP, dynamically deploying

[6] EGI FedCloud: https://www.egi.eu/services/cloud-compute/
[7] Several VREs have been created to provide users with data analytics facilities including:
https://services.d4science.org/group/envriplus (for everyone willing to experience with the facility),
https://services.d4science.org/group/eiscat (for supporting the IC_3 case),
https://services.d4science.org/group/icoseddycovarianceprocessing (for supporting the IC_13 case),
https://services.d4science.org/group/particleformation (for supporting the TC_17 case),
[8] DRIP: https://github.com/QCAPI-DRIP/

and managing as many Virtual Machines as required to cope with the load in order to be able to process the subscription requests in a timely manner.

1.5 Collaborations and Impacts

Apart from the technical contributions, agile teams largely support the collaboration across Work Packages, ENVRIplus RIs and organisations. Table 2 shows the involvement of WPs and ENVRIplus RIs in each use case, illustrating the collaborations established among the ENVRIplus community.

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

NO.	WPS	ENVRIPLUS RIS	ORGANISATIONS
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
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, IPSL, 52°NORTH, MARUM, CSEM, INRN
TC_16	WP5, WP9	DASSH, MEDIN	MBA, CU, UVA
TC_17	WP7, WP9	ICOS, SMEAR, ACTRIS, D4SCIENCE, EGI	UNIHB/PANGAEA, ISTI-CNR
SC_3	WP7, WP9	LIFEWATCH-SW, EGI, D4SCIENCE, BIOVEL	UGOT, EGI, ISTI-CNR

** SC: Science Case; TC: Test Case, IC: Implementation Case. Refer to D9.1 for detailed information.*

Using international conferences, workshops, and working groups (in various organisations and community networks) as vehicles, the agile teams have made broad impacts beyond ENVRI; a few selected examples are listed below:

- **RDA**: The Research Data Alliance (RDA) was launched as a community-driven organization in 2013. Today it has 7,000 members from 137 countries, with increasing global impact on building the social and technical infrastructure to enable open sharing of data. Individuals from several agile teams are active members of RDA: e.g., the IC_1 agile leader co-chairs the

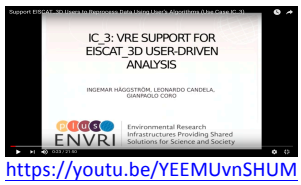
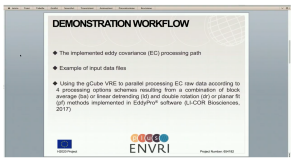
RDA Data Citation Working Group (DCWG), and IC_9 leader is a member of the research data collections working group[9].

- **EOSC**: The European Open Science Cloud (EOSC) opens up new era for European researchers and innovators to discover, access, use and reuse a broad spectrum of resources for advanced data-driven research. A number of agile teams (IC_1, IC_3 and TC_2) got involved as Competence Centers in the EOSC-hub project.
- **DI4R**: Digital Infrastructures for Research (DI4R)[10] is one of the high-impact conferences for e-Infrastructures, jointly organised by European leading e-Infrastructures and projects such as EOSC-hub, GEANT, PRACE, OpenAIRE, EGI, EUDAT etc. Two workshops proposed by WP9 were accepted and organised in the past DI4R conferences (2016 in Krakow, 2017 in Brussels). Various use cases (IC_1, IC_3, and TC_2) were presented as ENVRIplus project results.
- **EGU**: the European Geosciences Union (EGU) conferences are one of Europe's premier geosciences conferences attracting thousands of participants each year in Vienna. Agile teams such as IC_1, IC_9, IC_10, and TC_17 actively participate in the EGU network.

1.6 Final Science Demonstrators

Thanks to the agile leaders and their teams who have been actively testing Theme 2 service and various technologies, a subset[11] of use cases made excellent progress and are selected as final Science Demonstrators for Theme 2 services. By Theme 2 Science Demonstrator, we mean "a showcase of a service solution illustrated through a prototype implementation, which serves as proof or evidence that the Theme 2 services can bring added value for supporting the ENVRIplus community to conduct scientific research." This deliverable reports the final implementation results of those Science Demonstrators (listed in Table 3).

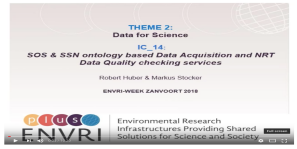
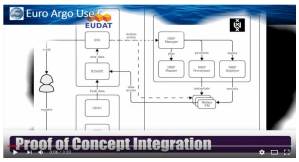
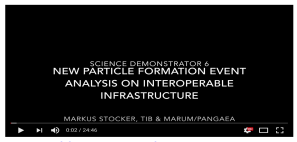
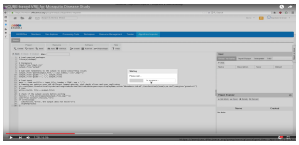
TABLE 3. LIST OF THE FINAL SCIENCE DEMONSTRATORS

#	SCIENCE DEMONSTRATOR	LINK TO THE DEMO
Science Demonstrator 1	**Title**: Support EISCAT_3D Users to Reprocess Data Using User's Algorithms (Use Case IC_3) **Author**: Ingemar Haggstrom, EISCAT Scientific Association, ingemar.haggstrom@eiscat.se	 https://youtu.be/YEEMUvnSHUM
Science Demonstrator 2	**Title**: The eddy covariance fluxes (Use Case IC_13) **Authors**: • Domenico Vitale, University of Tuscia, domvit@unitus.it • Dario Papale, University of Tuscia, darpap@unitus.it • Leonardo Candela, CNR-ISTI, leonardo.candela@isti.cnr.it	 https://youtu.be/hod2WksKzV8

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

[10] DI4R conferences https://www.digitalinfrastructures.eu/

[11] Other use cases that are not included in this report will be continue developed. And the results will be described in other reports or document.

	• Gianpaolo Coro, CNR-ISTI, gianpaolo.coro@isti.cnr.it	
Science Demonstrator 3	**Title**: SOS & SSN ontology based Data Acquisition and NRT Data Quality checking services (Use Case IC_14) **Author**: Rober Huber, University of Bremen, rhuber@uni-bremen.de	 https://youtu.be/p3UQZkRRWlw
Science Demonstrator 4	**Title**: EuroArgo Data subscription service (Use Case TC_2) **Authors**: • Thierry Carval, IFREMER, Thierry.Carval@ifremer.fr • Glenn Judeau, IFREMER, Glenn.Judeau@ifremer.fr • Jani Heikkinen, CSC, jani.heikkinen@csc.fi • Baptiste Grenier, EGI, baptiste.grenier@egi.eu • Zhiming Zhao, UvA, z.zhao@uva.nl • Paul Martin, UvA, p.w.martin@uva.nl • Spiros Koulouzis, UvA, S.Koulouzis@uva.nl	 https://youtu.be/PKU_JcmSskw
Science Demonstrator 5	**Title**: Sensor registry (Use Case TC_4) **Authors**: • Justin Buck, BODC, juck@bodc.ac.uk • Simon Jirka, 52°North, jirka@52north.org	 https://youtu.be/4QxTZ2iiznk
Science Demonstrator 6	**Title**: New particle formation event analysis on interoperable infrastructure (Use Case TC_17) **Authors**: • Markus Stocker, TIB and MARUM/PANGAEA, markus.stocker@tib.eu • Markus Fiebig, NILU, markus.fiebig@nilu.no • Leonardo Candela, CNR, leonardo.candela@isti.cnr.it • Giuseppe La Rocca, EGI, giuseppe.larocca@egi.eu • Enol Fernandez, EGI, enol.fernandez@egi.eu • Alex Hardisty, CU, hardistyar@cardiff.ac.uk	 https://youtu.be/ra9W7b5DbgI
Science Demonstrator 7	**Title**: gCube-based VRE for Mosquito Diseases Study (Use Case SC_2) **Authors**: • Baptiste Grenier, EGI, baptiste.grenier@egi.eu • Matthias Obst, Swedish Lifewatch, matthias.obst@marine.gu.se • Leonardo Candela, CNR-ISTI, leonardo.candela@isti.cnr.it • Gianpaolo Coro, CNR-ISTI, gianpaolo.coro@isti.cnr.it	 https://youtu.be/IBJkSys5tVo

The remainder of the document provides detailed information about the seven science demonstrators, compiled by the relevant agile teams.

2 Science Demonstrators

2.1 Science Demonstrator 1: Support EISCAT_3D Users to Reprocess Data Using User's Algorithms (Use Case IC_3)

Overview

Often the data products from a RI, derived from lower level/raw data, are pre-cooked. That means they are produced by applying some default processing algorithms and parameters, like spatial and temporal resolution, the use of model parameters, model and process selections. IC_3 demonstrates a model that enables scientific researchers to re-process data by defining other selections of these inputs from other sources.

Scientific Objectives

EISCAT_3D is an environmental research infrastructure on the European Strategy Forum on Research Infrastructures (ESFRI) roadmap. Once assembled, it will be a world-leading international research infrastructure to study the high 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 the original 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.

Description

A typical usage scenario is that users select data together with an algorithm and invoke a workflow 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 2 shows the EISCAT Real Time Graph (RTG), which is primarily developed to follow the radar run in real-time and to produce a display of basic parameters, such as returned power spectra.

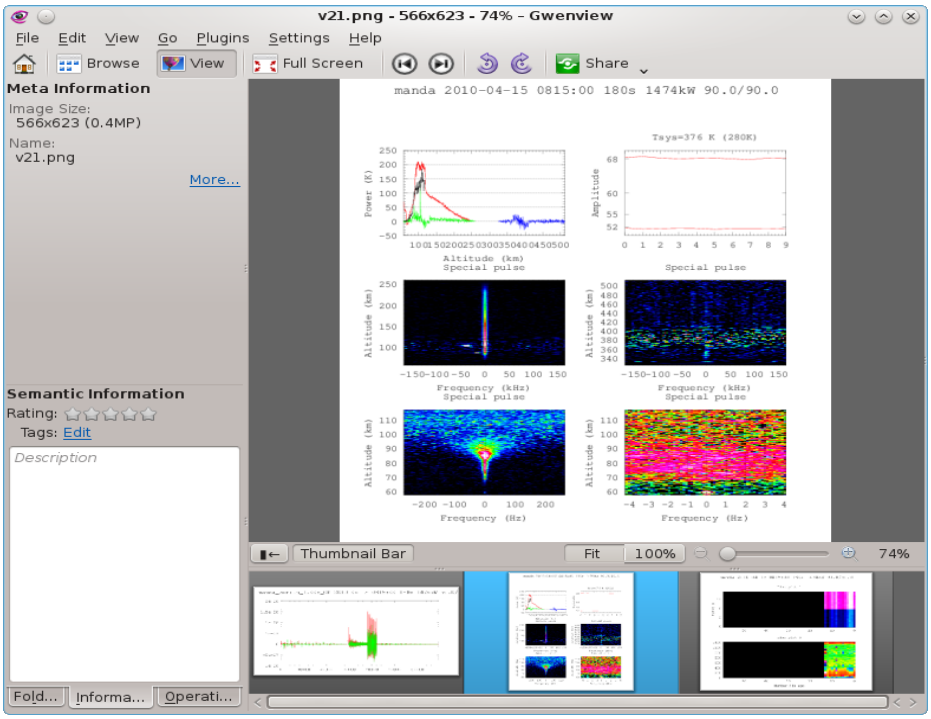


FIGURE 2. EISCAT REAL TIME GRAPH PRODUCES VISUALISATION OF THE PROCESSING OF RAW EISCAT RADAR DATA WITH BASIC PARAMETERS.

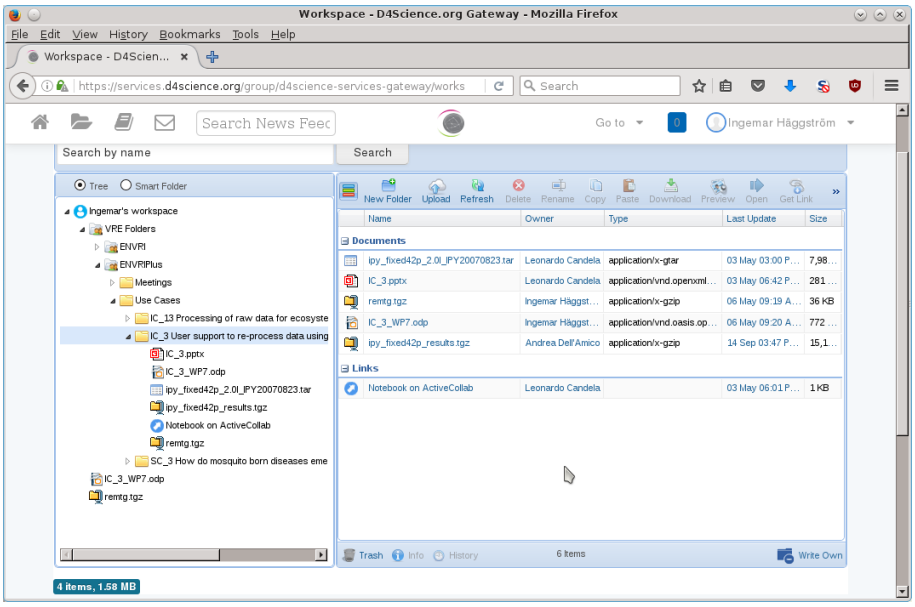


FIGURE 3. IN GCUBE, EISCAT USERS CAN CREATE PROCESSING ALGORITHMS AND SHARE THEM WITH OTHERS.

The RTG was modified to run under GNU Octave[12], and to produce the plots in batch mode with some selected input parameters. The whole setup was installed into the D4Science gCube Virtual Research Environment[13], provided by ENVRIplus WP7. The D4Science gCube Data Analytics is executed via a web dashboard based GUI. The ticketing system of D4science has been effectively

[12] GNU Octave: https://www.gnu.org/software/octave/
[13] The VRE is available at https://services.d4science.org/group/eiscat

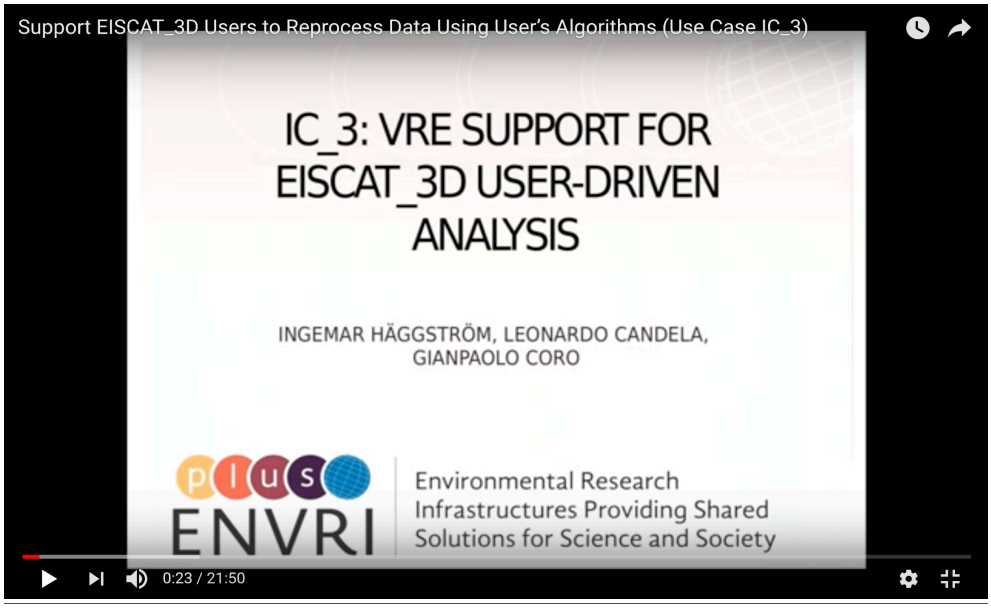
used for tracking the implementation and communicating with WP7 developers. Some initial results are shown in Figure 3, which demonstrates that in the gCube platform, the EISCAT RTG algorithms are translated into Octave, and compiled and published in an ENVRIplus community space. In the same way, an algorithm created by a single EISCAT user can be shared and reused by others.

Advantages

For many ENVRIplus RIs, the ultimate goal is to provide quality-checked observation data to their user communities. The ease of user access to data and analysis resources has a direct impact on the users' ability to deliver new research results and is thus directly linked with the impact and value of the RIs. This issue is therefore given increased emphasis when RIs design their ICT systems. In recent years, Virtual Research Environments (VREs) have emerged as an important approach to provide web-based systems helping researchers to collaborate. WP7 (T7.1) has set up a VRE for the ENVRIplus community using the D4Science platform. D4Science supports a flexible and agile application development model based on the notion of Platform as a Service (PaaS), in which components may be bound instantly at the time they are needed. In this way, it enables user communities to define their own research environments by selecting the constituents (the services, the data collections, the machines) among the pools of resources made available through the D4Science e-Infrastructure.

Access to a VRE for user-driven data analysis is a common request not only in the environmental sciences community. The implementations from this use cases can be easily extended to support other domain applications' needs, for example, it would benefit space and solar-terrestrial physics, solar system physics (meteors and asteroids), and astronomy. The pilot experiences can be easily shared with the whole EISCAT community, which covers member countries including, Finland, France, Norway, Sweden, USA and UK, and its global collaborations reach China, Japan, Korea.

Link to the Demonstrator



Link: https://youtu.be/YEEMUvnSHUM

The service prototype can be accessed at:

https://services.d4science.org/group/rprototypinglab/data-miner.

Simply request a use account, you will be able to login to the VRE. There, on 'Executes an experiment', selects ENVRI and Webtg V2.

Contributors

- Ingemar Häggström, EISCAT Scientific Association, ingemar.haggstrom@eiscat.se

2.2 Science Demonstrator 2: The Eddy Covariance Fluxes of GHGs (Use Case IC_13)

Overview

Eddy covariance (EC) fluxes calculation involves a complex set of data processing steps that, beyond the knowledge of the technique, requires a considerable amount of computational resources. This might constitute a constraint for RIs (e.g. ICOS) that aim to simultaneously process raw data sampled at multiple sites in Near Real Time (NRT) mode (i.e. provide each day fluxes estimates relative to the previous day). This demonstrator showcases the service solution that integrates the gCube service to optimize the processing of EC data based on 4 different processing schemes resulting from a combination of block average (ba) or linear detrending (ld) and double rotation (dr) or planar fit (pf) processing options and make this available to other RIs.

Scientific Objectives

The ambitious goal of this pilot investigation is to provide a computationally efficient tool able to process EC raw data and offer to users the possibility to calculate EC fluxes according to the multiple processing scheme described by Sabbatini et al. (2018)[14]. The ultimate aim is to establish a service that can be used by RIs that use this micrometeorological technique to measure exchanges of greenhouse gases and energy between terrestrial ecosystems and atmosphere (e.g. ICOS but also other RIs using the eddy covariance methods such for example LTER or ANAEE) .

Description

The EC technique involves high-frequency sampling (e.g. 10 or 20 Hz) of wind speed and scalar atmospheric concentration data, and yields vertical turbulent fluxes. EC fluxes are computed within a finite averaging time (normally 30 mins) from the covariance estimates between instantaneous deviations in vertical wind speed and gas concentration (e.g. CO_2) from their respective mean values, multiplied by the mean air density (see Aubinet et al., 2012)[15].

Despite the simplicity of this idea, a number of practical difficulties arise in transforming high-frequency data into reliable half-hourly flux measurements. To cope with these issues, here we used the tools implemented by the EddyPro® Fortran code (LI-COR Biosciences, 2017, Fratini and

[14] Sabbatini S. et al (2018). Eddy covariance raw data processing for CO_2 and energy fluxes calculation at ICOS ecosystem stations, International Agrophysics.

[15] Aubinet, M., Vesala, T., & Papale, D. (Eds.) (2012). *Eddy covariance: a practical guide to measurement and data analysis*. Springer Science & Business Media.

Mauder, 2014) [16] an open source software application avaialble for free download at https://www.licor.com/env/products/eddy_covariance/eddypro.html. The choice of EddyPro® software is motivated by *i*) the availability of different methods for data quality control and processing (e.g. coordinate rotation, time series detrending, time lag determination, spectral corrections, flux random uncertainty quantification, etc.), *ii*) the availability of the source code and *iii*) the fact that the software is based on a community developed set of tools.

Required for the processing of EC raw data through EddyPro® software, are 1) the availability of metadata information about the EC system setup and raw data file structure, and 2) the choice of a suitable combination of processing options.

Concerning 1), users have to provide a standardized metadata file in .csv format (metadata.csv, see Table 4). This file constitutes the input of an R script that automatically builds the mandatory files ingested into the EddyPro® software (i.e. the .metadata and .eddypro files) developed ad hoc for this exercise. The organization and name of the metadata variables is based on an international standard (BADM) used also in the USA network AmeriFlux. The format of the csv has been instead designed in order to develop a template easy to prepare by individual scientists and organized RIs. In case of NRT data processing, in order to perform part of the flux corrections (i.e. spectral corrections and planar fit), 5 additional configuration files are needed: planar_fit.txt, spectral_assessment_badr.txt, spectral_assessment_lddr.txt, spectral_assessment_bapf.txt, spectral_assessment_ldpf.txt. They can be obtained by specific EddyPro® runs based on long periods of data (at least one month of data is usually required for a consistent parameters estimation). The above files have to be placed together with EC raw-data files in an archive folder (data.zip) which will constitute the input file of the current implementation (see Figure 4).

TABLE 4. DESCRIPTION OF METADATA TO PROVIDE IN THE METADATA.CSV FILE

COLUMN	VARIABLE LABEL (FILE HEADER)	DESCRIPTION (UNITS)
1	SITEID	Official EC station code following the FLUXNET standards (CC-Xxx)
2	LATITUDE	Geographic latitude ([-90,90] from S to N, decimal)
3	LONGITUDE	Geographic longitude ([-180,180] from W to E, decimal)
4	ALTITUDE	Altitude of ecosystem under study (m)
5	CANOPY_HEIGHT	Distance between the ground and the top of the plant canopy (m)
6	SA_MANUFACTURER	Manufacturer of the sonic anemometer (currently only gill)
7	SA_MODEL	Model of the SA (currently only SA-Gill HS-50 or -100)
8	SA_SW_VERSION	Embedded software version of the SA
9	SA_WIND_DATA_FORMAT	The format of wind data (currently only uvw)
10	SA_NORTH_ALIGNEMENT	Specify whether the SA's axes are aligned to transducers (axis) or spars (spar)
11	SA_HEIGHT	The vertical distance between the ground and the center of the device sampling volume (m)
12	SA_NORTH_OFFSET	Specify the SA's yaw offset with respect to local magnetic

[16] Fratini, G., & Mauder, M. (2014). Towards a consistent eddy-covariance processing: an intercomparison of EddyPro and TK3. Atmospheric Measurement Techniques, 7(7), 2273-2281.

		north (degree positive eastward)
13	GA_MANUFACTURER	Manufacturer of the gas analyzer (currently only licor)
14	GA_MODEL	Model of the GA (currently only GA_CP-LI-COR LI7200)
15	GA_SW_VERSION	Embedded software version of the GA
16	GA_NORTHWARD_SEPARATION	The distance between the center of sample volume of the GA and the SA as measured horizontally along the north-south axis (cm)
17	GA_EASTWARD_SEPARATION	The distance between the center of sample volume of the GA and the SA as measured horizontally along the east-west axis (cm)
18	GA_VERTICAL_SEPARATION	The distance between the center of sample volume of the GA and the SA as measured along the vertical axis (cm)
19	GA_TUBE_DIAMETER	The inside diameter of the intake tube (mm)
20	GA_FLOWRATE	The flow rate of the intake tube (l/min)
21	GA_TUBE_LENGTH	The length of the intake tube (cm)
22	FILE_DURATION	The time span covered by each raw file (min)
23	ACQUISITION_FREQUENCY	The number of records per second in raw files (10 or 20 Hz)
24	FILE_FORMAT	Specify the format of raw files (ASCII or BIN)
25	FILE_EXTENSION	Specify the raw files extension (e.g. .csv, .txt, .dat)
26	LN	Logger number (from 1 to 10)
27	FN	Number of the file generated by the logger (from 1 to 10)
28	EXTERNAL_TIMESTAMP	0 or 1 if the timestamp in the file name refers to the beginning or the end of averaging period, respectively.
29	INTERNAL_TIMESTAMP	1 if there is a timestamp internal to raw files, otherwise 0.
30	EOL	Specify the end of line of raw files (e.g. lf)
31	SEPARATOR	The character that separates individual values in raw files
32	MISSING_DATA_STRING	Specify the character string used for missing data in raw files (e.g. NA, NaN, -9999)
33	NROW_HEADER	The number of rows in the header of the raw file
33+1	COLNAMES_1	Variable name in the first column of the raw data file
33+j	COLNAMES_j	Variable name in the j-th column of the raw data file
33+N	COLNAMES_N	Variable name in the last column of the raw data file

It is important to note that in the current implementation only few sensors are supported (the one used in ICOS) but the structure has been prepared in order to be ready to add new sensors and new processing methods, options and combinations.

The use of different processing options lead to different flux estimates. Discrepancies in flux estimates are caused by systematic errors introduced by methods used in the raw-data processing stage. Since there are not tools to establish a priori which is the best combination of processing options providing unbiased flux estimates, the viable solution, proposed by Sabbatini et al. (2018) and implemented here, involves a multiple processing scheme where EC flux data are calculated according to different combinations of methods.

In particular, , EC fluxes are calculated according to four different processing schemes resulting from a combination of block average (ba) or linear detrending (ld) and double rotation (dr) or

planar fit[17] (pf) processing options (for details see Aubinet et al, 2012)[18]. All other processing options remain unchanged: maximum cross-covariance method for time lag determination, spectral correction method proposed by Fratini et al. (2012)[19], statistical tests by Vickers and Mahrt (1997)[20] and by Foken and Wichura (1996)[21] for data quality control, method by Finkelstein and Sims (2001)[22] to estimate random uncertainty.

To reduce the computational runtime, the implementation of the four processing schemes above is performed in parallel mode in the gCube Virtual Research Environment (VRE). The processing path is defined as in Figure 4. When using EC raw data from a single observation tower, the estimated computational time required for a NRT run is about 4 minutes, similar to those required for the run of a single processing scheme.

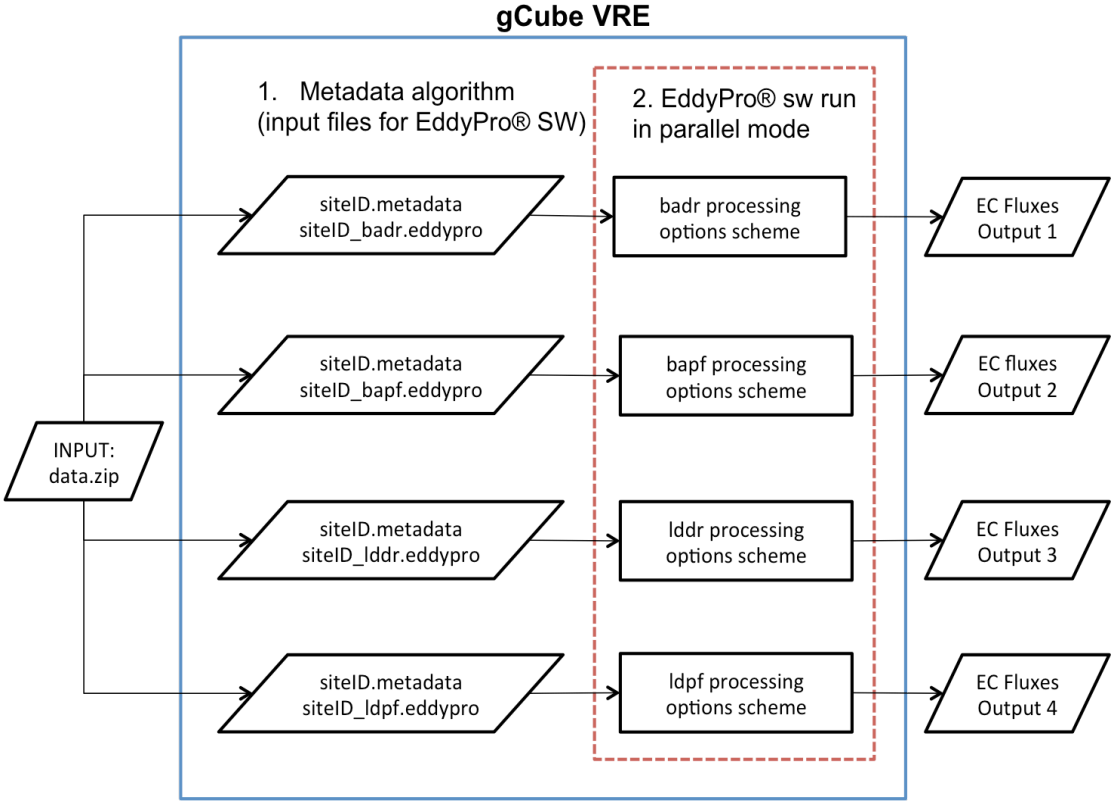


FIGURE 4. EC DATA PROCESSING PATH

[17] Wilczak, J. M., Oncley, S. P., & Stage, S. A. (2001). Sonic anemometer tilt correction algorithms. *Boundary-Layer Meteorology*, *99*(1), 127-150.

[18] Aubinet, M., Vesala, T., & Papale, D. (Eds.) (2012). *Eddy covariance: a practical guide to measurement and data analysis*. Springer Science & Business Media.

[19] Fratini, G., Ibrom, A., Arriga, N., Burba, G., & Papale, D. (2012). Relative humidity effects on water vapour fluxes measured with closed-path eddy-covariance systems with short sampling lines. Agricultural and forest meteorology, 165, 53-63.

[20] Vickers, D., & Mahrt, L. (1997). Quality control and flux sampling problems for tower and aircraft data. Journal of Atmospheric and Oceanic Technology, 14(3), 512-526.

[21] Foken, T., & Wichura, B. (1996). Tools for quality assessment of surface-based flux measurements. Agricultural and forest meteorology, 78(1-2), 83-105.

[22] Finkelstein, P. L., & Sims, P. F. (2001). Sampling error in eddy correlation flux measurements. Journal of Geophysical Research: Atmospheres, 106(D4), 3503-3509.

Advantages

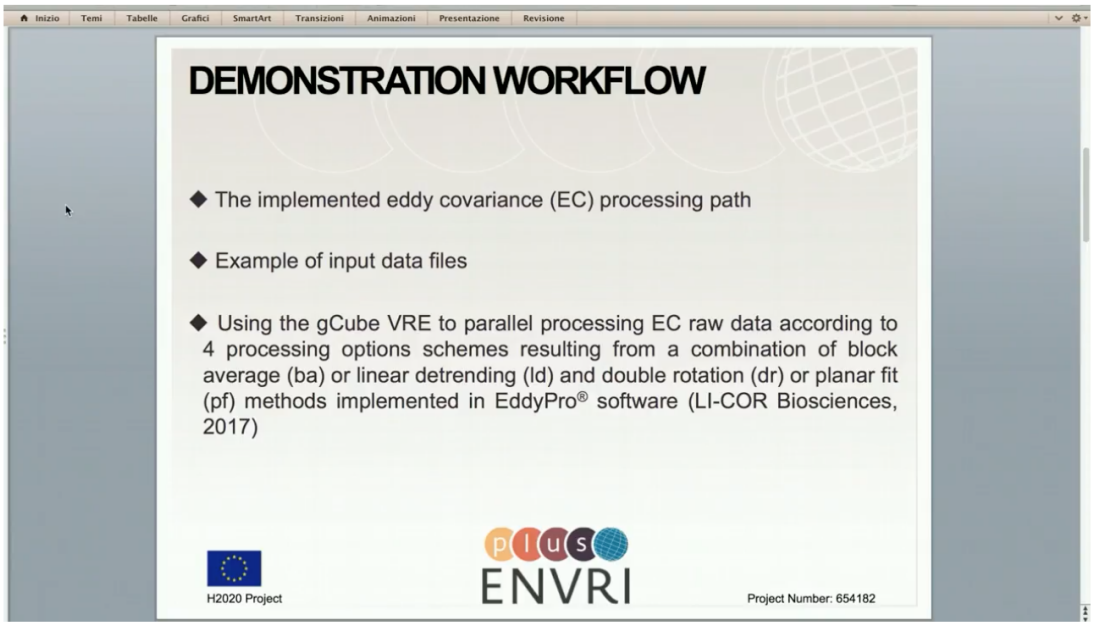
The implementation of a multiple processing schemes as illustrated above and the direct management and use of metadata according to international standard in the eddy covariance community constitutes a novelty in the context of EC data analysis. The main advantage of the multiple processing is twofold. From one hand, it offers the possibility of an extensive evaluation of the effect each method has on flux data estimation. On the other, by combining the output results as described by Sabbatini et al. (2018), it is possible to obtain more consistent estimates of the uncertainty associated to EC fluxes.. The direct use of metadata instead ensure the needed flexibility for a large use of the tool if the new sensors are added in the system.

The efficiency of parallel computing implemented in the VRE, drastically reduces the computational runtime required to obtain flux estimates from different processing options schemes. This constitutes a clear advantage for any user and in particular, for RIs aiming at analyzing routinely large amount of data.

Although, here we selected only 4 processing option schemes, the efficiency of parallel computing implemented in the VRE offers the possibility to increase the number of processing schemes suitable for the EC data processing and also post-processing steps.

This might considerably improve our understanding about the performance of methods developed for EC raw-data processing and about interpretation of resulting fluxes.

Link to the Demonstrator



Link: https://youtu.be/hod2WksKzV8

Contributors

- Domenico Vitale, University of Tuscia, domvit@unitus.it
- Dario Papale, University of Tuscia, darpap@unitus.it
- Leonardo Candela, CNR-ISTI, leonardo.candela@isti.cnr.it
- Gianpaolo Coro, CNR-ISTI, gianpaolo.coro@isti.cnr.it

2.3 Science Demonstrator 3: SOS & SSN Ontology Based Data Acquisition & Near Real Time Quality Control (Use Case IC_14)

Overview

The Service allows to submit and publish raw observational (non-geophysical) environmental timeseries data in common standard formats (T-SOS XML and SSNO JSON) via a messaging API (EGI ARGO[23]) that is used to perform Near Real Time (NRT) quality control procedures by an Apache Storm NRT QC Topology, which publishes the quality controlled and labelled data via a messaging output queue.

Scientific 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 often 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).

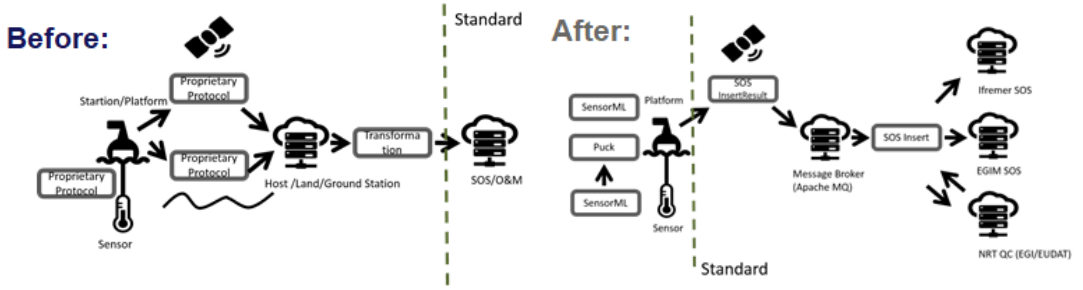


FIGURE 5. MOVING STANDARDS CLOSER TO THE SENSOR. TODAY MANY RIS USE PROPRIETARY PROTOCOLS AND FORMATS WITHIN THEIR DATA FLOW (BEFORE). STANDARDS SUCH AS SSNO (NOT SHOWN HERE), SOS (SENSOR OBSERVATION SERVICE) OR O&M (OBSERVATION & MEASUREMENT) FREQUENTLY ARE ONLY USED TO PUBLISH DATA. OUR APPROACH WAS TO USE THESE STANDARDS AS EARLY AS POSSIBLE (AFTER): USE PUCK TO EXPOSE SENSORML METADATA, TRANSMIT DATA AS T-SOS INSERTRESULT XML AND ISSUE THESE DATA AT POTENTIALLY MULTIPLE SOS SERVERS AS WELL AS FOR CONSUMPTION BY A NRT QC SERVICE.

A further objective is to contribute to the harmonization of data transmission formats and protocols.

This science demonstrator use case aims to evaluate standardized data transmission using OGC SWE Transactional SOS (Sensor Observation Service) as a priority standard as well as using the Semantic Sensor Network (SSN) ontology. Both are implemented and tested. It will identify and implement common generic NRT QC routines suitable for multiple RIs (e.g. EMSO, EuroARGO, ANAEE, etc.) and deploy these on appropriate scalable cloud based technologies at own and/or EGI platforms.

[23] EGI ARGO: https://wiki.egi.eu/wiki/ARGO, http://argo.egi.eu/

Description

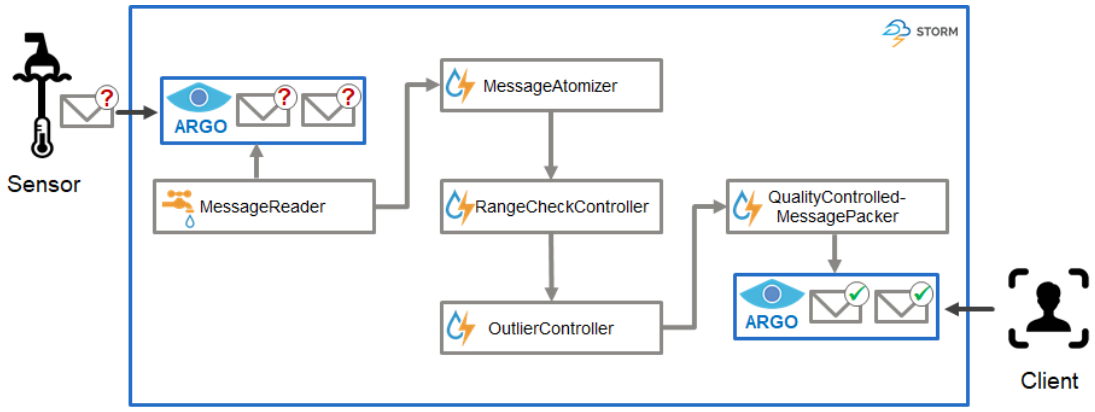


FIGURE 6. ARCHITECTURE OF THE NRT QC SERVICE

The service is based on two main cloud based components: 1) an Apache Storm data processing unit responsible for near real time data quality control of a given real-time time series data and 2) a EGI ARGO messaging component responsible for the management and queuing of data sent from the sensor as well as for the delivery of quality controlled data.

As illustrated in Figure 6, the data processing and quality control builds on Apache Storm to support scalable NRT QC on streamed standardized sensor data. Apache Storm[24] 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. Central to Apache Storm is the notion of Storm topology. Topology 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.

Data from a sensor system using transactional SOS typically is sent in distinct intervals. We therefore chose to use EGI's ARGO messaging system which is built upon Kafka to manage such incoming messages. ARGO provides a well-documented REST API which is easy to use and based on JSON envelopes for messages wherein base64[25] encoded content can be sent. We have set up one message queue (topic) for incoming messages and another queue for delivery of quality labelled data.

Supported data messages are either base64 encoded T-SOS XML strings or SSNO based JSON strings. In order to ease things for processing we decided to transform these data messages into individual atomic observation objects based on SSNO.

Within the Storm topology we have implemented several nodes:

MessageReader is a BaseRichSpout which continuously reads messages from the ARGO message queue. As ARGO can send multiple messages within one API response, the spout splits theses messages into individual message objects and emits these as tuples for further processing within the topology.

[24] Apache Storm: http://storm.apache.org/
[25] base64 is a group of binary-to-text encoding schemes that represent binary data in an ASCII string format by translating it into a radix-64 representation.

MessageAtomizer Bolt is a BaseRichBolt which collects sensor metadata from sensor URLs given in the T-SOS or SSNO such as sensor specific measurement ranges. It recognizes the sent data format (SSNO or T-SOS), splits the message objects into atomic observation object tuples and emits these tuples.

RangeCheckController Bolt is a BaseRichBolt which takes these emitted tuples and checks if each numeric value is within the measurement range of the sensor specification. It adds a qualityOfObservation value (0=passed, 1=failed) to each atomic value and emits these as tuples.

OutlierController Bolt is a BaseWindowedBolt which collects a given number of atomic value tuples and performs a simple outlier check based on a modified z-score. The bolt again adds a qualityOfObservation value to each checked atomic value and emits these as tuples.

QualityControlledMessagePacker Bolt is a BaseWindowedBolt which collects a given number of quality checked atomic values and adds these into a JSON array which is the sent as payload to the ARGO messaging queue for QC'd messages where it is available for further consumption by an appropriate domain specific service to update or clean raw data holdings.

Advantages

Near real time quality control is a common problem for Research Infrastructures. Whereas domain specific NRT routines and standards exist, such as those for ICOS, ARGO or IAGOS, we have shown (see ENVRIplus deliverable D3.3[26]) that clear communalities among those RIs with respect to NRT QC routines exist.

.	EMSO	FixO3	EuroARGO	ACTRIS (surf.)	ACTRIS (lidar)	IAGOS	ICOS (ETC)	ICOS (OTC)	GROOM		Count
Data Integrity Test	-	-	-	x	-	-	X	-	-		1
Metadata Consistency Test	-	-	-	x	x	x	X	-	-		4
Platform/Sensor Identification	x	x	x	x	-	-	X	-	x		8
Date/Time Check	x	x	x	-	-	x	X	x	x		7
Location Check	-	x	x	-	-	x	-	x	x		5
Spike or Outlier Test	x	x	x	x	-	x	x	x	x		8
Gradient Test	x	x	x	-	-	-	-	-	x		4
Stuck or Constant Value Test	-	x	x	-	-	x	X	x	x		6
Digit Rollover Test	-	-	x	-	-	-	-	-	x		2
Gap Test	-	-	-	-	-	x	x	-	-		2
Rate of Change / Step Test	-	x	-	-	-	x	X	-	-		3
Range Test (Regionality/Past Values)	-	x	x	-	-	x	-	x	x		5
Range Test (Global)	-	x	x	-	-	x	X	-	-		4
Range Test (Instrument Limits)	-	-	-	x	-	x	x	x	-		4
Range Test (Implicit)	-	-	-	x	x	x	-	-	-		3

FIGURE 7. COMMONLY USED NRT QUALITY ROUTINES WITHIN ENVRIPLUS RIS

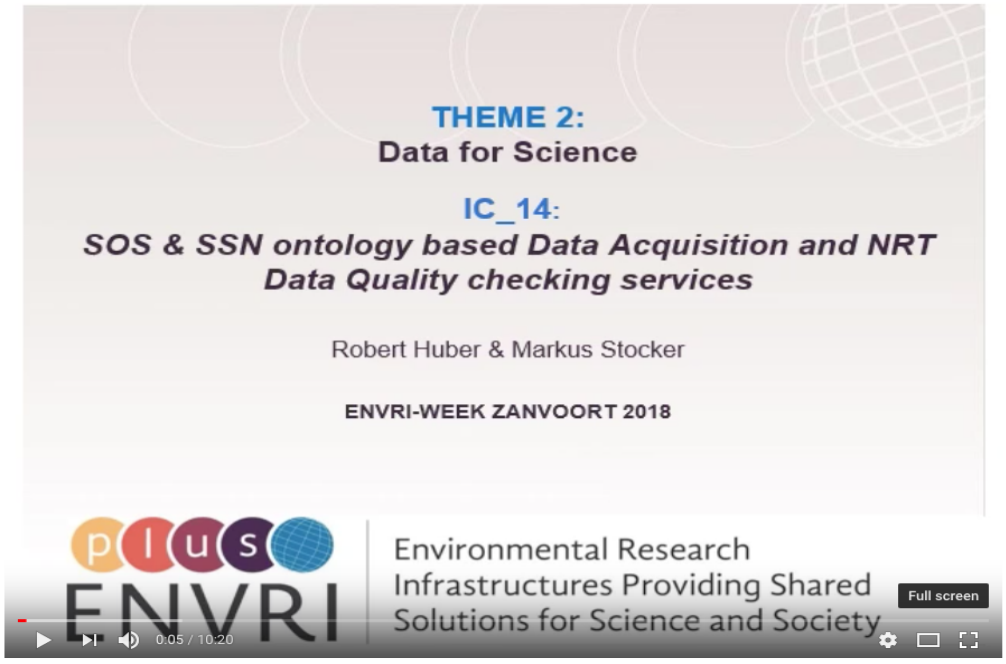
Most commonly used are simple test such as outlier or spike detection, gradient or stuck value tests. A domain independent service able to perform these tests would therefore be an added value for all ENVRIplus RIs. RIs which have their own services in place could use it for cross-validation of their QC results and it would give those RIs the opportunity to perform routine NRT QC checks which do not yet have own routines in place.

[26] ENVRIplus deliverable D3.3: http://www.envriplus.eu/wp-content/uploads/2015/08/D3.3.pdf

Unfortunately, data transmission formats used within RIs is very diverse. In general, data transmission at the sensor as well as platform level largely depends on community specific needs and habits or simply on manufacturer specifications. It is therefore difficult to offer generic, cross RI processing services in general and in particular services which allow NRT QC. It is therefore clearly advantageous for the scientific community to have access to standard supporting services. Further, such services potentially can strongly promote the use of these standards - one of the main objectives of this use case.

Link to the Demonstrator

- Video



Link: https://youtu.be/p3UQZkRRWlw

- Github repository: https://github.com/ab-e/lightning/tree/demo

Contributors

- Rober Huber, University of Bremen, rhuber@uni-bremen.de

2.4 Science Demonstrator 4: EuroArgo Data Subscription Service (Use Case TC_2)

Overview

The EuroArgo Data Subscription Service (DSS) allows researchers to subscribe to customized views on Argo data, selecting specific regions and time-spans, and choosing the frequency of updates. Tailored updates are then provided on schedule to researchers' private storage.

Scientific 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 at the Argo data center 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. 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 use 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 service. The new developed service allows data registration through data identification. Optionally the community registers the subscription actions and their parameters in B2SAFE. Data could be requested through interfaces provided by EUDAT and IFREMER web sites. Subscription is managed by EUDAT and actions are processed using EGI FedCloud. Users can select different attributes for their subscription like localization, stations, parameter, update frequency and more.

Description

Architecture

As shown in (Figure 8), the Data Subscription Service (DSS) involves the following basic components: 1) a data selection portal as frontend, 2) the Global Data Assembly Center (GDAC) of EuroArgo, 3) EUDAT B2SAFE storage, 4) DRIP, 5) EGI FedCloud resources, and 6) a subscription service component for managing the subscriptions registered via the data selection portal.

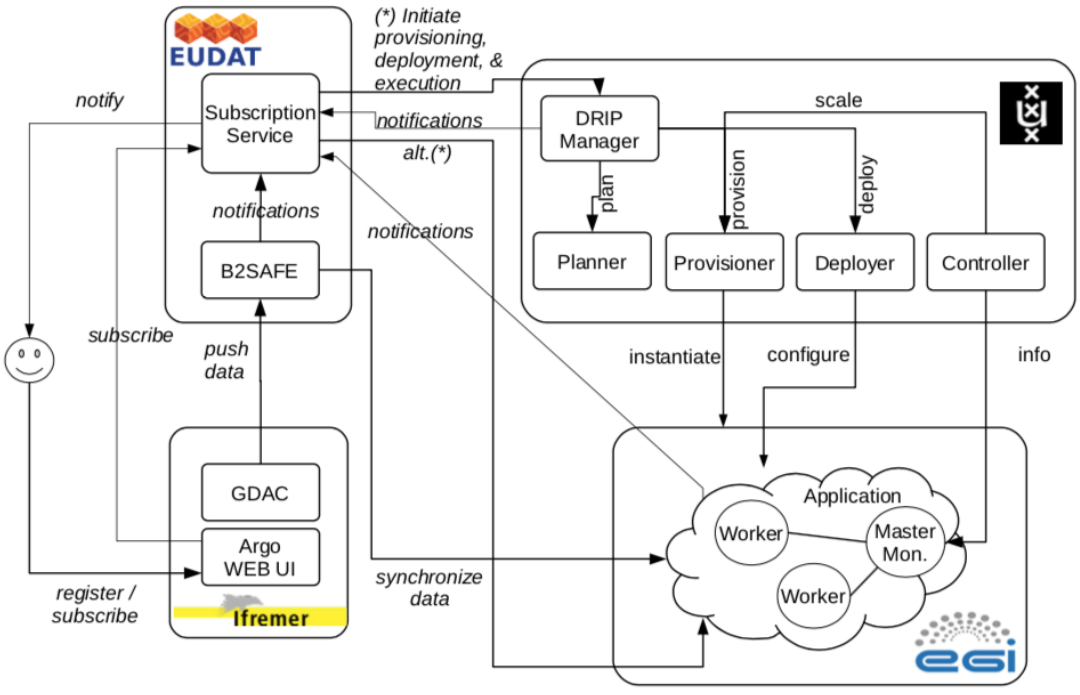


FIGURE 8. SYSTEM ARCHITECTURE FOR EUROARGO DATA SUBSCRIPTION SERVICE.

IFREMER provided accumulated monthly marine domain datasets to be pushed to EUDAT e-Infrastructure (B2SAFE). The datasets were then synchronized between EUDAT and EGI Cloud resources on demand. EUDAT provides services for storage and data transfer, while EGI FedCloud provides the services for computing data products for each subscription.

A Web Portal has been developed for users to select and subscribe to the interested 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 with user's criteria.

DRIP (the Dynamic Real-time Infrastructure Planner, developed by WP7, Task T7.2) is integrated to execute the data selection process using parallel computation. DRIP can dynamically deploy and manage 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 and the user was notified by email.

Workflow

The typical workflow is as follows: users interact with the DSS via the portal, registering to receive updates for specific areas and time ranges for selected parameters such as temperature, salinity, and oxygen levels. The GDAC receives new datasets from regional centres and pushes them to the B2SAFE data service. The DSS maintains records of subscriptions including selected parameters and associated actions. DRIP plans, provisions, deploys, scales and controls the data filtering application. EGI FedCloud provides cloud resources to host the application. The application itself is composed of a master node and a set of worker nodes.

When new data is available to the GDAC, it pushes them to the B2SAFE service, triggering a notification to the DSS, which consequently initiates actions on the new data. If the application is not deployed to FedCloud then DRIP provisions the necessary VMs and network so that the application may be deployed. Next, the deployment agent installs all the necessary dependencies along with the application including configurations to access the Argo data. The DSS signals to the application master node the availability of the input parameters to be processed, whereupon it partitions the input tasks into sub-tasks and distributes them to the workers. If the input parameters include deadlines then the master will prioritise them accordingly. The monitoring process keeps track of each running task and passes that information to the DRIP controller. If the programmed threshold is passed, then the controller will request more resources from the provisioner. Finally, the results of each task are pushed back to the B2SAFE service triggering a notification to the subscription service, after which it notifies the user.

User Interface

Shown in Figure 9, users can use DSS web portal to subscribe to interested datasets. A typical subscription task is made up of a set of inputs: a) an area expressed as a bounding box; b) a time range; c) a list of parameters required in data products (e.g. temperature); and d) optionally, a deadline.

FIGURE 9. DSS WEB PORTAL

Advantages

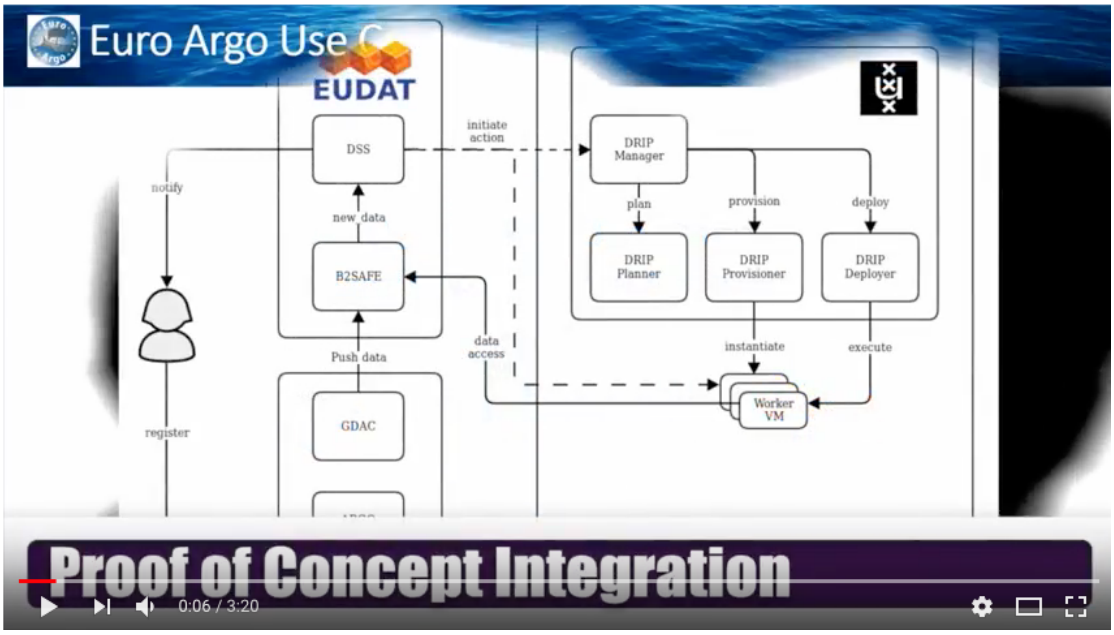
The pilot activity was initiated by the marine research community, however, the possibility to receive regular transmissions of data, especially in near-real time, directly from the organisation responsible for the data collection and (pre-)processing, is very important to many large initiatives. Generic initiatives will themselves be interested to operate subscription services for their outputs, based around a trusted repository hosting synchronised versions of their data collections. Such a service may also allow the provision of new features to end users, generating more visibility.

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

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

Data subscription services are expected to play an increasing role in the future, as the number of data producers and their respective output continues to increase rapidly. The mechanism tested and implemented by this demonstrator could contribute to the development of both a common standard for input streams to enhancements of digital collaborative spaces for researchers and data providers.

Link to the Demonstrator



Link: https://youtu.be/PKU_JcmSskw

Contributors

- Thierry Carval, IFREMER, Thierry.Carval@ifremer.fr
- Glenn Judeau, IFREMER, Glenn.Judeau@ifremer.fr
- Jani Heikkinen, CSC, jani.heikkinen@csc.fi
- Baptiste Grenier, EGI, baptiste.grenier@egi.eu
- Zhiming Zhao, UvA, z.zhao@uva.nl
- Paul Martin, UvA, p.w.martin@uva.nl
- Spiros Koulouzis, UvA, S.Koulouzis@uva.nl

2.5 Science Demonstrator 5: Sensor Registry (Use Case TC_4)

Overview

The “sensor registry” aims at supporting the management of sensors deployed for in-situ measurements. Common sensors or families of sensors are used across different research infrastructures, for example, oxygen optodes that are equipped on platforms in multiple research infrastructures. The goal of this work is to define common methods to access the sensor metadata in such cases.

Four sub-use cases are considered:

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

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[27]) as well as solid earth (e.g., EPOS) or atmosphere observations (e.g., ICOS).

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

Scientific Objectives

The objective of Standardised sensor repositories to enable:

- Easily discover sensors and their metadata
- Sensors and sensor observations discoverable, accessible and usable via the web via standardised services
- Seamlessly integrate sensors from one network with sensors from other networks

The sensor registry is based on OGC Sensor Web Enablement technologies (SWE) that have been developed and implemented by a range of national and European projects and activities. As part of the FP7 and H2020 projects ODIP, ODIP2, Oceans of tomorrow call and the BRIDGES project the Marine SWE profile was developed. The marine SWE profile is a marine specialisation of the OGC SensorML (for metadata) and OGC Observations & Measurements (O&M) standards. A significant advance within these templates was the use of controlled vocabularies to describe terms and values within the SensorML documents. The NERC vocabulary server (NVS, part of the European SeaDataNet infrastructure) was used to host the vocabularies. The vocaularies are the Wxx family of vocabularies and accessible at https://www.bodc.ac.uk/resources/vocabularies/vocabulary_search/. In addition to the human interface the NVS has machine readable access methods including SPARQL and SOAP with outputs including RDF, XML and JSON. The standards developed in the OoT projects are summarised in the SenseOcean project joint deliverable D7.8[28] which also included future recommendations for future work.

Key outcomes of the marine SWE profile are:

- Standardized web services will exist for accessing sensor information and sensor observations
- Sensor systems will be capable of real-time mining of observations to find phenomena of immediate interest (event stream processing)
- ISO/OGC O&M as data format and the OGC SOS as data access interface enable EC INSPIRE directive compliance with the data format
- The associated OGC Sensor Observation Service (SOS) would be the natural way to achieve full INSPIRE compliance (there exists additional INSPIRE technical guidance that describes how to use the SOS as INSPIRE download service).

Our goal in this test case is to demonstrate which elements of the standardised interface to data are available and how they could potentially be integrated. The development has happened

[27] EuroGOOS Gliders: http://eurogoos.eu/gliders-task-team/
[28] SenseOcean project deliverable D7.8: http://www.senseocean.eu/senseocean/sites/senseocean/files/documents/Deliverable%20D7.8%20Policy%20Document%20Sensor%20Development%20for%20the%20Ocean%20of%20Tomorrow_r.pdf

across a number of projects and the alignment of the results is not fully achieved because each project has different requirements and deliverables. Consequently examples will be shown from different activities demonstrating each goal in the introduction. The outcomes of the WP9 Test Case TC_4 Sensor registry include:

- Proposed goal to attempt to link repositories to users for specific examples
- Demonstrate viability of the technology and test interoperability
- Act as a precursor to future projects with broader linkages and harvesting

Description

1. List and discover specifications of sensors and hardware on the market.

For the RIs 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.

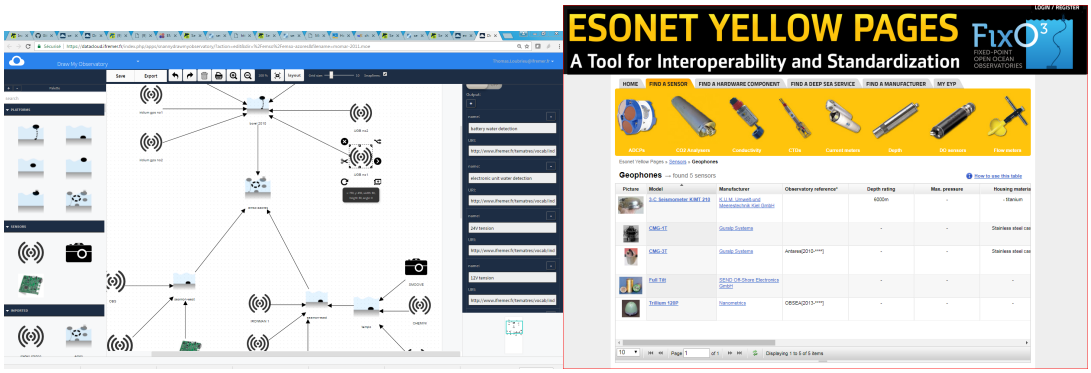


FIGURE 10. SENSOR DEPLOYMENT GRAPHICAL EDITOR AND SENSOR MODEL DATABASE FOR EMSO

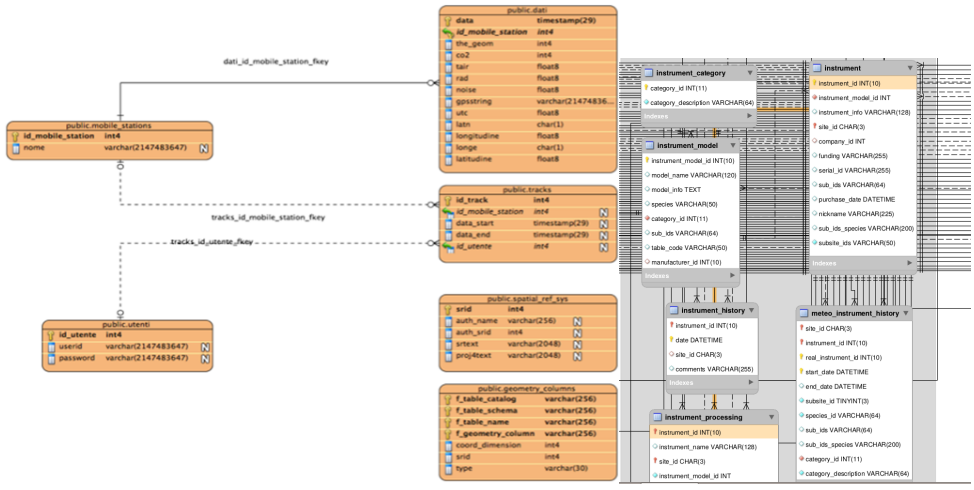


FIGURE 11. DATA MODEL FOR ANAEE AND ICOS SENSOR DESCRIPTIONS

Regarding standards, a sample of SIOS station description has been encoded in OGC/SensorML format. Examples of SSN ontology encoding and translation from SensorML to SSN have been provided from FixO3 project inputs.

One of the commonly used software packages for providing OGC SWE services in European projects is the 52°North Sensor Observation Service (SOS). This provides an interoperable web-

based interface for inserting and querying sensor data and sensor descriptions. It aggregates observations from live in-situ sensors as well as historical data sets (time series data). This server component is complemented by an extensible Web-based JavaScript Sensor Web viewer which allows the visualisation of different types of sensor observation data: the 52°North Helgoland Viewer. More information is available from: https://52north.org/ .

These concepts have been used by the Oceanids Command and Control project along with the outcomes of the Marine SWE profile to store and expose the metadata for ocean glider deployments to the web. These metadata are then used to automate the curation and conversion of raw glider data to the EGO NetCDF (http://archimer.ifremer.fr/doc/00239/34980/) format which is the data exchange format within the Ocean Glider Network. Oceanids uses a database that supports both SSN and SensorML metadata representations so is strongly aligned with this demonstrator. The same database is being used for historical EMSO data (specifically the PAP site) that is held by BODC and development is on-going to expose the metadata in SensorML and data in O&M formats, this is scheduled to be complete in late 2018.

2. Manage the park of sensor by owner, manage maintenance (e.g. calibrations), loan, …

The recording of a sensor history is now technically possible and facilitated by the SensorML template produced within the Marine SWE profile. To date this has not been fully implemented to the authors knowledge. BODC will be introducing the linking of documentation to SensorML records in the development scheduled prior to March 2019.

3. Edit deployments, enable traceability from observation data back to the sensor and procedures used for acquisition (link with implementation case on provenance IC_3).

For users of the 52°North software (one of the primary open source OGC SWE software suites) there has been a SensorML editor SMLE (pronounced smilee) that enables users to interact with and edit SensorML documents held in OGC compliant SOS servers via a graphical user interface (editing is only supported for SOS servers supporting the transactional SOS operations). This is freely available from: https://github.com/52North/smle

4. Discover infrastructures (observation network, equipped experiment sites...) and enable their citation.

The discovery of infrastructures requires the recently developed federated sensor observation services. These have not been trialed or investigated on the marine domain to the authors knowledge at the time of writing. They may need refinement for the specific requirements of the use case, akin to the Marine SWE profile activity. Work on evaluation of these services was a key recommendation of the OoT projects (see previous reference above).

The citation of sensors is a on-going development activity in a newly formed Research Data Alliance working group on the persistent identification of sensors (https://www.rd-alliance.org/groups/persistent-identification-instruments-wg).

Advantages

Standardised machine-readable sensor metadata links directly to the catalogue and provenance work in WP8 of ENVRIplus, it provides a machine-readable document can be linked to data filling a current capability gap in the provenance trace.

Common metadata standards allow the sharing of metadata and federation between RIs, especially as the scientific community moves toward observation requirements that span multiple RIs

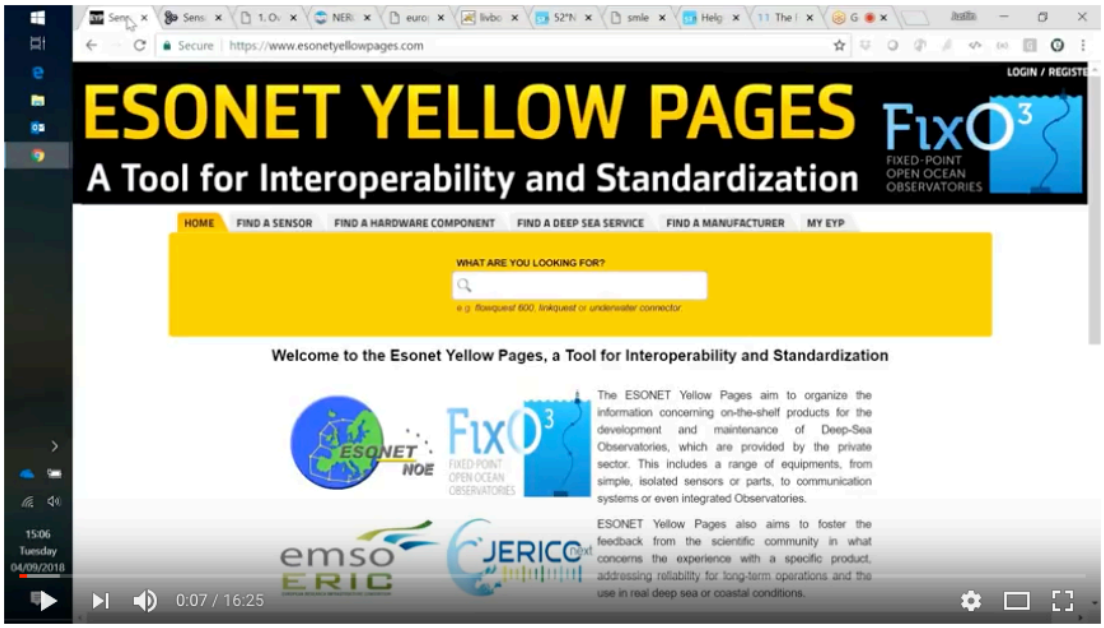
The service has potential users spanning the data lifecycle and value chain. The standards are applied at the time of observation and can be embedded within sensors. At the other end of the chain they will enable users to dynamically harvest and discover data for use cases such as oil spill response. Using standard such as OGC SOS and ISO/OGC O&M this will even result in the INSPIRE compliant provision of observation data.

Usage scenarios include:

- Embed in the standards within sensors or data processing systems to allow the automated processing and dissemination of data
- Standardised machine readable recording and linking of metadata to a data provenance trace
- Inclusion of data in OGC SWE services and endpoints
- The standardised sharing of sensor metadata between organisations and RIs
- The citation of specific sensors (dependent on the results of the RDA working group)

Link to the Demonstrator

Video



Link: https://youtu.be/4QxTZ2iiznk

Link to ESONET yellow pages: https://www.esonetyellowpages.com/

Link to selected 52North instances:

- 52°North SOS server used for publishing data collected in the NeXOS project: http://nexos.demo.52north.org/52n-sos-nexos-test/sos?request=GetCapabilities&service=SOS

- NeXOS Sensor Web Viewer based on the 52°North Helgoland client: http://nexos.demo.52north.org/client/

Example SensorML output from that follows the Marine SWE profile:

- BODC
  - A model of an Aanderaa oxygen optode: http://linkedsystems.uk/system/prototype/TOOL0969 /current/
  - An instance of an oxygen optode: http://linkedsystems.uk/system/instance/TOOL0969_prospect/current/
- OGS
  - An instance of a Wind Monitor-JR: http://europa.ogs.trieste.it/OGS_SOS/SensorML_3_0/Sensor_V3_E2M3A_WIND.xml
  - An instance of SBE 37-SMP-ODO MicroCAT high-accuracy conductivity and temperature recorder: http://europa.ogs.trieste.it/OGS_SOS/SensorML_3_0/Sensor_V3_E2M3A_CT.xml

Contributors

- Justin Buck, BODC, juck@bodc.ac.uk
- Simon Jirka, 52°North, jirka@52north.org

2.6 Science Demonstrator 6: New particle formation event analysis on interoperable infrastructure (Use Case TC_17)

Overview

For the scientific community in aerosol sciences that studies atmospheric new particle formation events (NPFEs), this service aims to prototype how the scientific community can be deeply integrated with interoperable Research Infrastructures and e-Infrastructures (unless specified otherwise henceforth referred to as infrastructures). The result is a knowledge infrastructure[29] i.e., a robust network of scientists, artefacts such as virtual research environments and research data, and institutions such as research infrastructures and e-Infrastructures that acquire, maintain and share scientific knowledge about the natural world.

The service demonstrates how data analysis can be exposed to researchers as a Web based service while interoperable infrastructures orchestrate everything else, specifically: (1) loading primary observational data into computing environments for subsequent analysis by researchers; (2) representation of data derived in analysis using data models that employ domain-relevant community vocabularies and capture machine readable data semantics i.e., information[30] ("meaningful data"); (3) systematic and automated acquisition of derivative information in infrastructures; (4) registration of derivative information in catalogues.

Scientific Objectives

[29] Edwards, Paul N. 2010. A Vast Machine: Computer Models, Climate Data, and the Politics of Global Warming. MIT Press.
[30] Floridi, L.: The Philosophy of Information. Oxford University Press (2011)

This demonstrator aims to prototype the future of scientific data analysis on interoperable infrastructure. It showcases how well-engineered infrastructures can provide to science communities data analysis as a service while taking care of everything else e.g., data conversions, curating data derived in analysis - as described in this section.

Here, the scientific community focuses on what they are most interested in and most enjoy doing (i.e., address scientific hypotheses with data analysis and interpretation) and the infrastructure guarantees that their data are FAIR[31] i.e., findable and accessible by systematic and automated acquisition and cataloguing; interoperable by using a formal, accessible, shared, and broadly applicable language for knowledge representation and vocabularies that follow FAIR principles; and reusable by rich description of data using domain-relevant community vocabularies and by their release with a clear data usage license.

An important objective is to entirely erase the need for manual data download and upload by researchers. The download of data from research infrastructures is "considered harmful" in most cases[32]. Indeed, the practice of downloading data perpetuates the infrastructural discontinuity between local computing environments (e.g., researchers' workstations) and engineered infrastructures. Such discontinuity makes it difficult or impossible for engineered infrastructures to monitor workflows and executed activities, retain information about the involved primary and derivative data, as well as to systematically acquire derivative data. We want to demonstrate what a knowledge infrastructure may look like when manual download and upload is not an option.

A second objective is to unravel what occurs in the data use phase of the research data lifecycle. Studied for a concrete use case in aerosol science involving infrastructures and the relevant scientific community, we analyse the details of a scientific data analysis workflow and the roles of both the scientific community and infrastructures as elements of knowledge infrastructures. The demonstrator showcases how primary observational data acquired, curated and published by a research infrastructure are analysed by the scientific community in the data use phase and how such analysis generates derivative data. Traditionally, derivative data are poorly standardized in the community and generally reside on the workstations of researchers. The demonstrator shows how, when data analysis is performed on interoperable infrastructures - thus avoiding the aforementioned infrastructural disconnect - infrastructures can guarantee systematic and automated acquisition of derivative data, thus ensuring a strong link between the data use phase and the (derivative) data acquisition phase of the research data lifecycle. The demonstrator emphasises that there indeed is a cycle for research data from primary data to scientific knowledge communicated in scholarly literature.

A third objective is to connect, i.e., deeply integrate, a scientific community with well-engineered infrastructures. In the future, this social objective needs to be given more emphasis. While there continue to be issues to iron out, the technical elements of knowledge infrastructures are today mature enough to move into full-scale operation. The experiences collected with this use case

[31] Wilkinson, M.D. et al.: The FAIR Guiding Principles for scientific data management and stewardship. Scientific Data 3 (mar 2016). https://doi.org/10.1038/sdata.2016.18

[32] Atkinson, M., Filgueira, R., Spinuso, A., Trani, L.: Download considered harmful (2018), manuscript in preparation.

underscore that the social elements are lagging behind and are at best only slowly starting to be receptive to the novel approaches developed here, at least in the earth and environmental sciences and especially for the long tail of science i.e., for communities that do “small science” with “little data”. This demonstrator is a contribution to integrating the social and technical elements of knowledge infrastructures.

Description

Usage Scenarios

Jaana and Mikko are two fresh graduate students of a Finnish research group that, among other things, study new particle formation events. New particle formation events are atmospheric events whereby aerosol particles form and grow over the course of a day at specific spatial locations. These events are studied to increase our understanding for the formation process and to quantify the formed aerosol.

Prior to Jaana and Mikko, earlier generations of students have developed Python software and published the codes as a Jupyter Notebook on GitHub. These codes have become the *de facto* standard software used by the scientific community. Jaana and Mikko are instructed by their supervisor to use these codes for their analysis and have obtained further instructions on how to execute the analysis on e-Infrastructures by their postdoc colleagues.

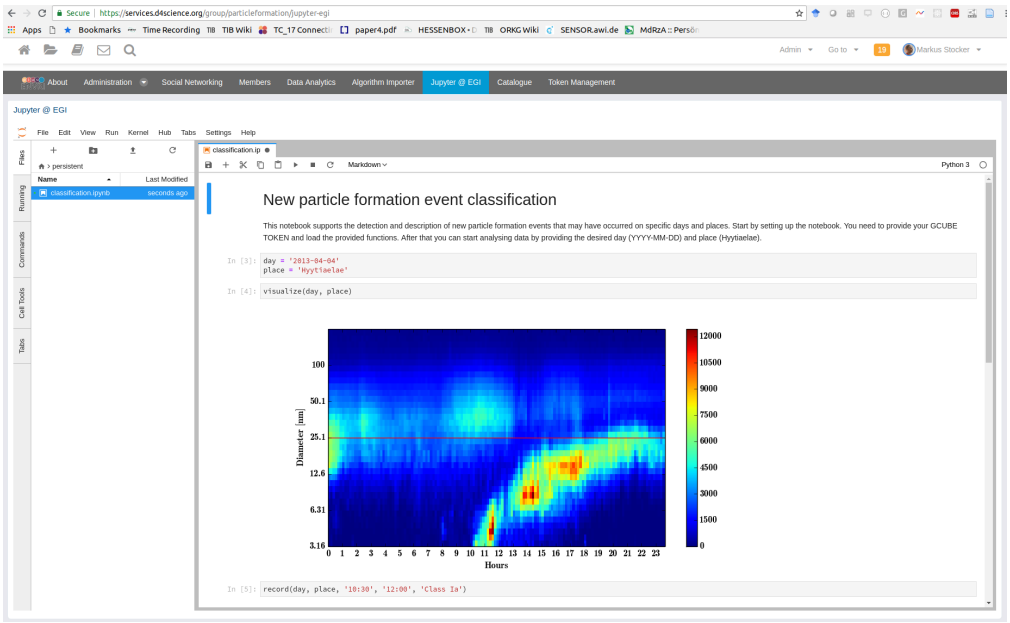


FIGURE 12. JUPYTER NOTEBOOK FOR NEW PARTICLE FORMATION EVENT CLASSIFICATION AS SEEN FROM THE D4SCIENCE VRE. THE VRE INCLUDES THE EGI JUPYTERLAB COMPUTING ENVIRONMENT.

The two graduate students create an account on D4Science and are given access to the relevant Virtual Research Environment (VRE). The VRE gives them access to JupyterLab, serviced by EGI. The students use the JupyterLab Terminal to clone the required Jupyter Notebook from GitHub into their own working space. Now the students are ready to analyse primary data (i.e., particle size distribution data) in order to detect and describe new particle formation events that may have occurred at specific places and days. Figure 12 displays the Jupyter Notebook as seen by Jaana and Mikko. All they need to do is to select a day and place and interpret the corresponding

visualization of primary data. For days and places at which an event occurred, the result of primary data interpretation is a description of the event, recording in particular the beginning and end times as well as the classification of the event, which follows a scheme accepted by the relevant scientific community.

Architecture

Figure 13 provides an overview of the architectural design of the service implementation. Researchers access JupyterLab operated on the EGI e-Infrastructure (provided by WP9) in order to analyse primary data for the purpose of new particle formation event detection and description. JupyterLab is accessible from the corresponding D4Science Virtual Research Environment[33] (VRE). Having cloned the required Jupyter Notebook[34] from GitHub, researchers can start to analyse primary data to detect and describe new particle formation events.

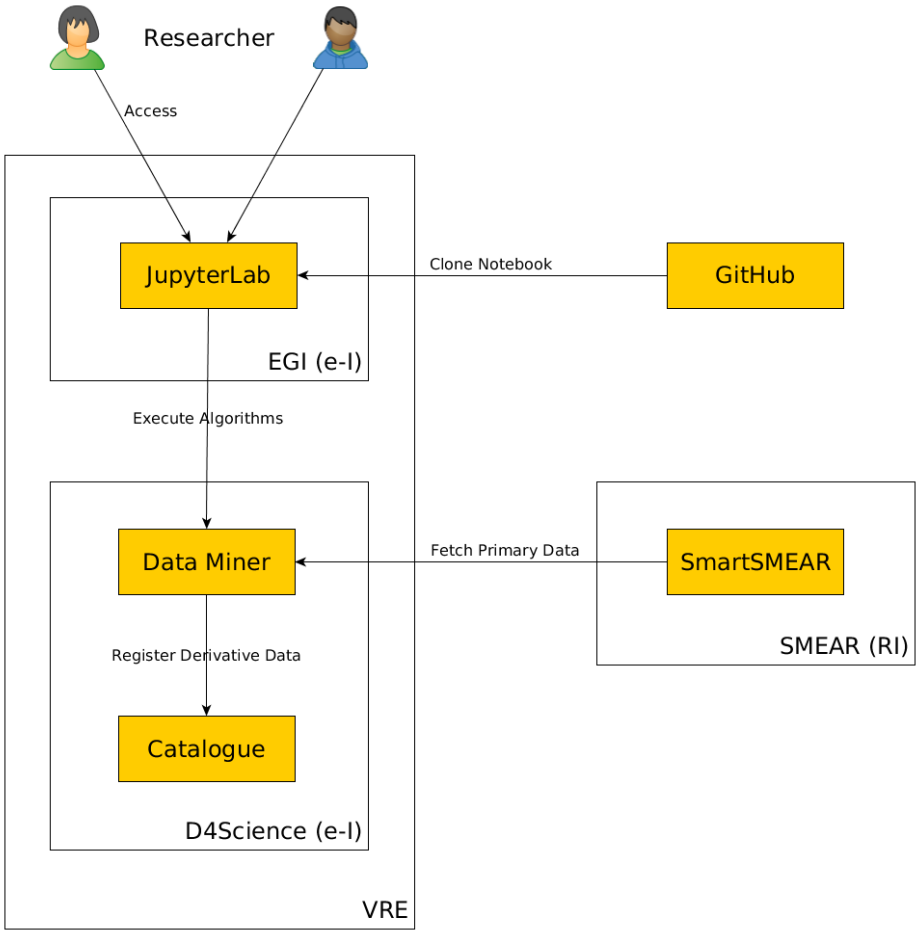


FIGURE 13. ARCHITECTURAL DESIGN OF THE SERVICE IMPLEMENTATION.

Workflow and Interfaces

The analysis consists of two main steps. Both are implemented as D4Science Data Miner algorithms and are accessed from within the Jupyter Notebook, programmatically via a WPS (OGC Web Process Service) interface. Given a day and place, as configured by the researcher, the

[33] https://services.d4science.org/group/particleformation/
[34] https://github.com/markusstocker/pynpf-d4science/blob/master/classification.ipynb

first step fetches and visualizes primary data. The primary data are published by SmartSMEAR[35], a "data visualization and download tool for the database of continuous atmospheric, flux, soil, tree physiological and water quality measurements at SMEAR research stations of the University of Helsinki." SmartSMEAR is developed and provided in collaboration with CSC (https://www.csc.fi/home), the Finnish national supercomputing center, who also host the SMEAR data. SmartSMEAR is thus an (software) artifact of the SMEAR[36] (Station for Measuring Ecosystem-Atmosphere Relations) research infrastructure (RI). SmartSMEAR provides an API for data access. The primary data can thus be fetched and loaded into Python data structures in a programmatic manner.

Given the primary data, the Data Miner algorithm creates a visual representation (see Jupyter Notebook in Figure 12). This visualization is used by researchers to decide whether a new particle formation event occurred on the selected day and place as well as to describe the event for its properties e.g., beginning and end times, classification, among others. The visualization is a conventional PNG image which is made accessible by D4Science. The Data Miner algorithm returns to Jupyter Notebook the URL for the location of the image. The notebook then visualizes the image by retrieving it from the given location.

Assuming an event occurred on the selected day and place, the result of interpreting the visualization is a description of the event. Since researchers here study new particle formation events, in this context an event description is information i.e., meaningful data. Such data are thus rich in semantics.

In the second step, an additional Data Miner algorithm records the event description. The researcher merely records the day (e.g., 2013-04-04), place (e.g., Hyytiälä), beginning (e.g., 11:00), end (e.g., 12:30) and classification (e.g., Class Ia). Rather than recording these strings into a row of a table, the algorithm creates an RDF (Resource Description Framework) description of the event. The meaning of the strings is thus recorded as well. The result is a self-describing information object (see Figure 15). The current implementation uses the LODE[37] ontology, which provides a concept Event and relations for time and space. We are currently working with the scientific community to develop a more appropriate concept of new particle formation events[38]. This concept will be part of the Environment Ontology[39]. When this process is completed, we will modify the implementation to reflect the conceptualization developed by the scientific community.

Finally, the RDF description is registered as a resource on the CKAN based D4Science catalogue. This is done automatically by the Data Miner algorithm. The data derived in analysis are thus automatically catalogued with corresponding metadata that support search. Figure 14 shows a number of catalogued resources for Hyytiälä (FI) on various days. Figure 15 shows the metadata of a selected resource, specifically the one for the description of the new particle formation event that occurred at Hyytiälä on April 4, 2013. Here, users are also given the location (URL) from which the event description can be accessed. Finally, Figure 16 shows the RDF description

[35] https://avaa.tdata.fi/web/smart
[36] http://www.atm.helsinki.fi/SMEAR/
[37] http://linkedevents.org/ontology/
[38] https://github.com/EnvironmentOntology/envo/issues/602
[39] http://www.obofoundry.org/ontology/envo.html

(in Turtle syntax) of the new particle formation event that occurred at Hyytiälä on April 4, 2013, obtained by accessing the URL on D4Science.

The RDF description is machine-readable, uses formal languages for knowledge representation (RDF, RDFS, OWL), and represents the semantics of the data (the day, place, beginning, end and classification) derived in analysis using domain-relevant community vocabularies. Specifically, beginning and end data elements (10:30 and 12:00, respectively) are described as an OWL-Time[40] Interval with beginning and end OWL-Time Instants, relating to the respective timestamps as XSD DateTime. Also notable is that the place is identified as a GeoNames[41] resource (http://sws.geonames.org/656888/) for Hyytiälä, Finland. The data derived in analysis are thus richly annotated. Indeed, the description adopts Linked Data principles to relate to resources defined elsewhere (here, GeoNames.org).

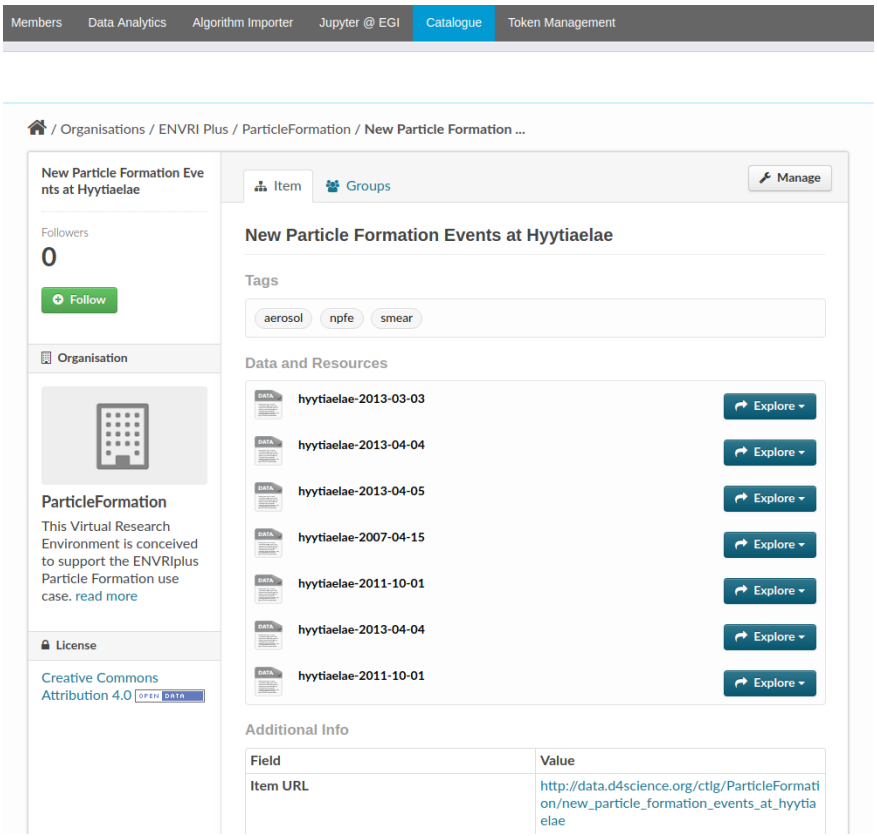


FIGURE 14. NEW PARTICLE FORMATION EVENT DESCRIPTIONS REGISTERED AS RESOURCES IN THE CKAN BASED D4SCIENCE CATALOGUE.

[40] https://www.w3.org/TR/owl-time/
[41] http://www.geonames.org/

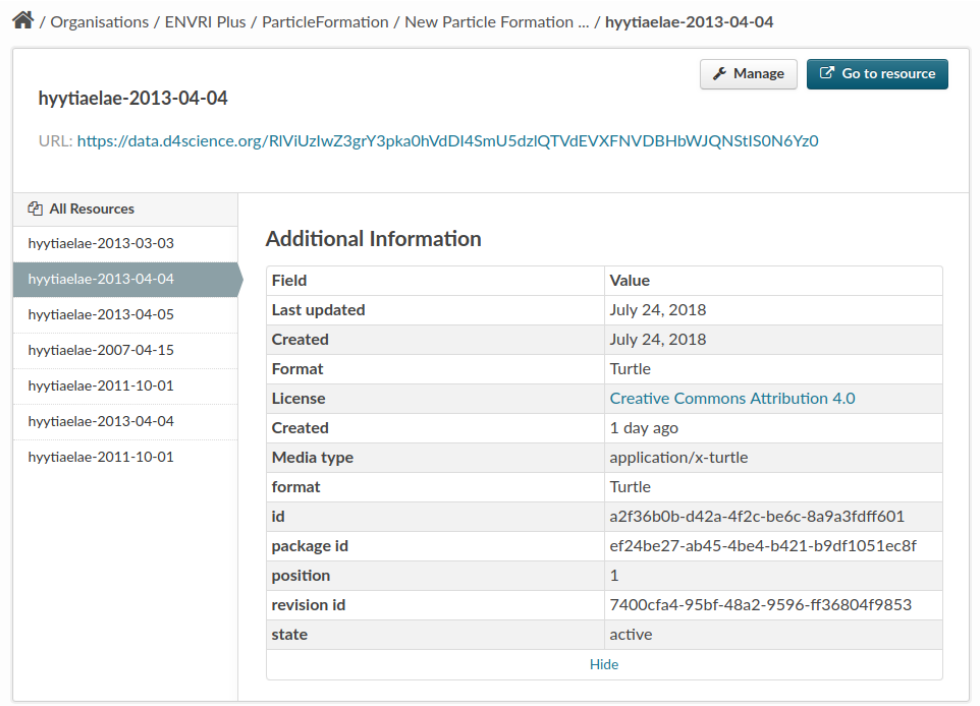


FIGURE 15. METADATA AND ACCESS URL OF THE RESOURCE DESCRIBING THE NEW PARTICLE FORMATION EVENT THAT OCCURRED AT HYYTIALA (FI) ON APRIL 4, 2013.

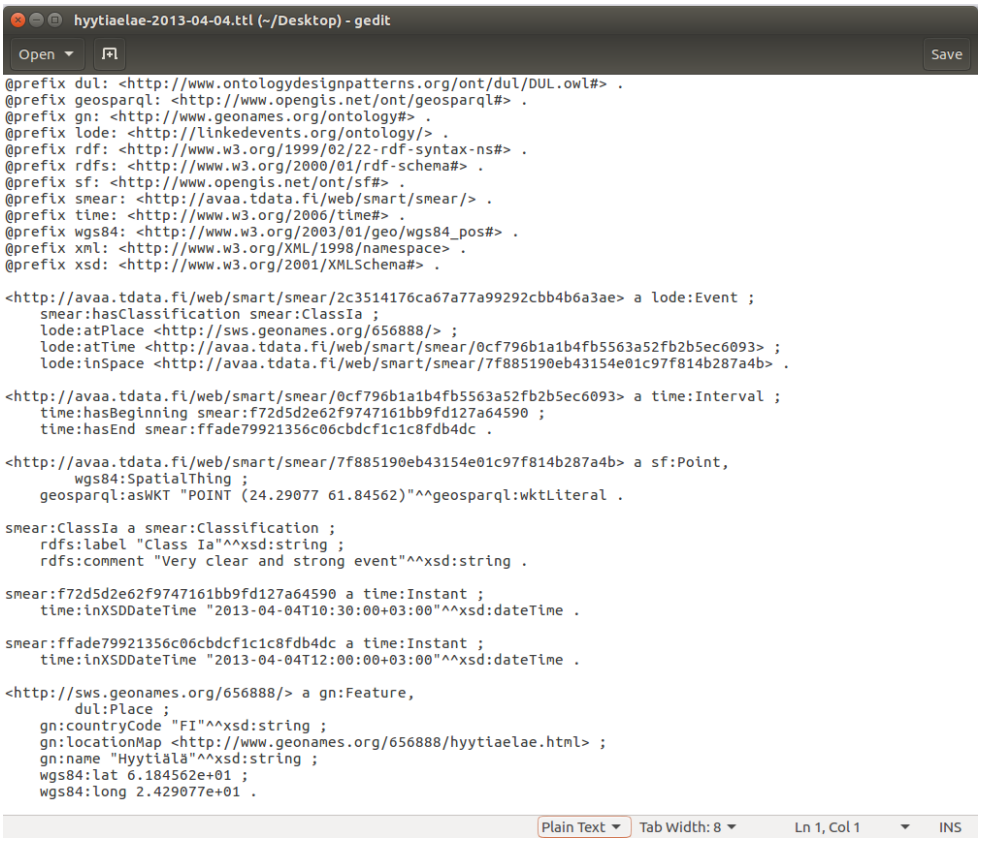


FIGURE 16. THE RDF DESCRIPTION OF THE NEW PARTICLE FORMATION EVENT THAT OCCURRED AT HYYTIALA ON APRIL 4, 2013 AS RETRIEVED FROM THE D4SCIENCE CATALOGUE.

Advantages

The idea of transforming data into knowledge is popular among research infrastructures. Among others, the Integrated Carbon Observation System (ICOS) research infrastructure uses the tagline "knowledge through observations"[42]. The European Multidisciplinary Seafloor and water column Observatory (EMSO) suggests that the research infrastructure plays "a major role in supporting the European marine sciences and technology [...] to enter a new paradigm of knowledge in the XXI Century"[43]. As an example beyond research infrastructures, the European Open Science Cloud (EOSC) is envisioned to be an environment that enables turning ever increasing amounts of data "into knowledge as renewable, sustainable fuel for innovation in turn to meet global challenges"[44].

Beyond the specifics of the developed use case in aerosol science, this demonstrator is a clear contribution to this idea. It demonstrates a possible architecture of an infrastructure that “transforms data into knowledge”. Essential factors of such knowledge infrastructures are (1) the deep integration of science communities with research and e-Infrastructures; and, as an important technical factor, (2) the curation of formal (i.e., machine-readable) data semantics. The deep integration of science communities is essential because, never mind the Age of Artificial Intelligence, in science it is researchers that transform data into knowledge. As this demonstrator underscores, deep integration with infrastructures allows for a range of novel possibilities, in particular enable researchers to focus on data analysis and interpretation while leaving data access and transformation from and to systems, the representation of data and their semantics following community standards, the capture of provenance information, and other infrastructural aspects to infrastructures.

The curation of data semantics is an additional essential, technical, factor. Information is inherently semantic and becomes knowledge through learning (i.e., internalization). When researchers interpret primary data, the resulting derivative *information* is thus rich in meaning (relative to the context of data analysis). As demonstrated here, the data tuple (2013-04-04, Hyytiälä, 11:00, 12:30, Class Ia) has a specific semantic. In the Age of Semantics, it is paramount to move on from data structures that merely capture such a tuple of values to data models that also support the machine readable representation of the meaning of these values. We understand that it can’t be expected from researchers to (manually or, to this date also, programmatically) on broad disciplinary scale translate data tuples such as the one above into a description as shown in Figure 16. Indeed, a key point of this demonstrator is to showcase the possibility of engineering such translation into infrastructures so that the whole is entirely invisible to researchers. Such invisibility reflects well the nature of infrastructure[45].

Analysing 17 ENVRIplus research infrastructures, Hardisty et al.[46] have reported that at the time (2016) only three identify core competencies in the data use phase of the research lifecycle. As

[42] https://twitter.com/ICOS_RI/status/803156982349729793
[43] http://www.emsodev.eu/work_packages.html
[44] https://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf
[45] Star, S.L.: The ethnography of infrastructure. American Behavioral Scientist 43(3), 377–391 (1999). https://doi.org/10.1177/00027649921955326
[46] Hardisty, A. et al. (2016). A definition of the ENVRIplus Reference Model. ENVRIplus Deliverable 5.2. http://www.envriplus.eu/wp-content/uploads/2015/08/D5.2-A-definition-of-the-ENVRIplus-Reference-Model.pdf

per the definition of the ENVRI Reference Model[47], data use is the phase in which researchers use data, potentially producing new research data. This description best describes the activity supported by this demonstrator, which thus focuses on the data use phase. While only few ENVRIplus research infrastructures identify core competencies in the data use phase, it is surely correct that all research infrastructures serve the data use phase. Indeed, their very existence is to serve this phase, serve researchers' data use. A key kind of use is arguably data analysis and interpretation. Again, beyond the specifics of the developed use case in aerosol science, we argue that this demonstrator showcases one possible interplay between research infrastructures and e-Infrastructures in the data use phase. We argue that the developed architecture is applicable to other (ENVRIplus) research infrastructures and the science communities they serve. Hence, the principles developed by this demonstrator are broadly applicable. Naturally, the specifics (e.g., the vocabularies, the notebook, the cataloguing, etc.) need to be adapted to meet the requirements of other scientific communities. The architecture and implementation principles (e.g., the design and technologies used) are, however, broadly applicable and thus of relevance to other or probably most if not all (ENVRIplus) research infrastructures.

Since FAIR data is on the global agenda of infrastructures, funders and other institutions, we underscore that this demonstrator significantly contributes to implementing this agenda by promoting the notion of "FAIR by Design" - weaving data FAIRness into infrastructures' fabric. The demonstrator builds on the principle not to leave making data FAIR to researchers but to guarantee it by design of well-engineered infrastructures. We argue that the removal of manual download and upload of data from and to systems is a crucial factor to this effect.

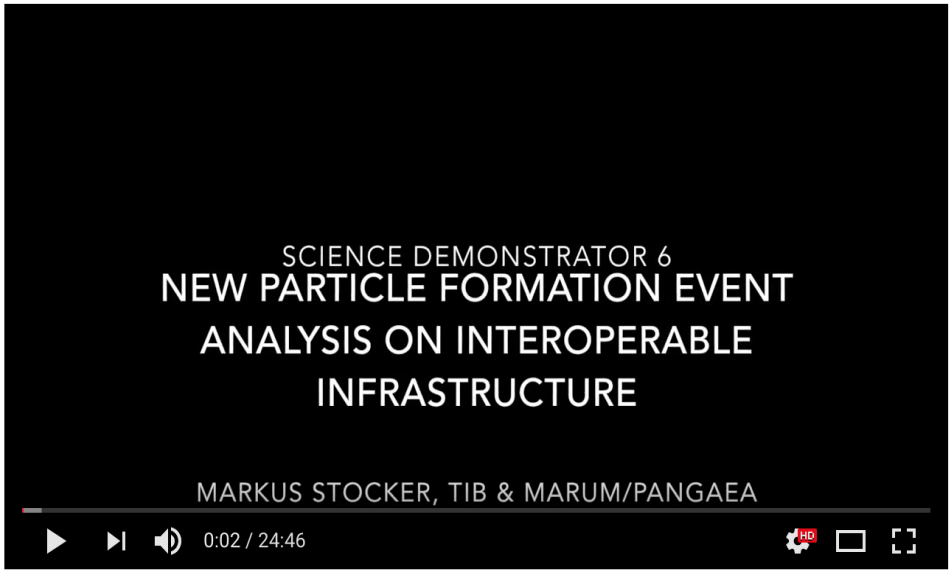
Naturally, the demonstrator is first and foremost of primary interest to a specific scientific community, namely the one consisting of the various aerosol research groups that study new particle formation events. To the best of our knowledge, the globally most renown research group in this area is the one led by Prof. Markku Kulmala at University of Helsinki[48]. Prof. Kulmala and some of the postdocs in his group have been involved in the developments of this demonstrator. Most importantly, postdocs have been actively involved in the development of a conceptualization of new particle formation events and a corresponding concept of the Environment Ontology. Naturally, in its current stage the demonstrator is a prototype to showcase to the scientific and infrastructure communities what is possible using state of the art interoperable infrastructures. A transition in practice from how data analysis is currently done to such infrastructures as demonstrated here requires further work as well as further acceptance by the scientific community. While we think to have reached an important milestone with this demonstrator, we cannot claim to know if and when such a transition will occur, for this scientific community or beyond. Clearer is, however, the imperative of the transition toward a practice as delineated by this demonstrator.

Link to the Demonstrator

- Instructions: https://github.com/markusstocker/pynpf-d4science/blob/master/README.md
- Virtual Research Environment https://services.d4science.org/group/particleformation/

[47] https://envri.eu/rm
[48] https://www.helsinki.fi/en/inar-institute-for-atmospheric-and-earth-system-research

- Blog post
  http://markusstocker.com/data-analysis-on-interoperable-infrastructure/
- Video



Link: https://youtu.be/ra9W7b5DbgI

Contributors

- Markus Stocker, TIB and MARUM/PANGAEA, markus.stocker@tib.eu
- Markus Fiebig, NILU, markus.fiebig@nilu.no
- Leonardo Candela, CNR, leonardo.candela@isti.cnr.it
- Giuseppe La Rocca, EGI, giuseppe.larocca@egi.eu
- Enol Fernandez, EGI, enol.fernandez@egi.eu
- Alex Hardisty, CU, hardistyar@cardiff.ac.uk

2.7 Science Demonstrator 7: gCube-based VRE for Mosquito Diseases Study (Use Case SC_2)

Overview

This demonstration illustrates how a LifeWatch researcher can easily upload and integrate an R-based algorithm in D4science, making it available to other researches, in particular 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 processing data-intensive analysis algorithms, the computation can be outsourced on federated resources, such as those provided by the EGI e-Infrastructures.

Scientific 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.

LifeWatch RI provides advanced ICT, such as BioVel, supporting biodiversity research. However, it currently only provides standard algorithms for data processing. There is a need to support

individual researchers' requests, e.g., import a new set of hydrological data layers into the analysis, add new algorithms that handle presence/absence into analysis etc., and a need for access to Cloud resources, e.g., to execute a large number of analytical cycles for many species under different climate scenarios.

These 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.

Description

Architecture

The proposed service architecture is shown in Figure 17. It combines different infrastructures: at a lower layer is the LifeWatch RI, containing the Swedish LifeWatch Portal that provides high-quality biological data for mosquito species, and the community data repositories that preserves environmental information and a series of ecological modelling algorithms. Datasets to be exploited include species data (95,730 abundance measurements from Sweden, Denmark, and Germany for 40 disease-carrying species in 2016), and hydrological data (generated by a regional hydrological model using 15 land use types and 8 soil types).

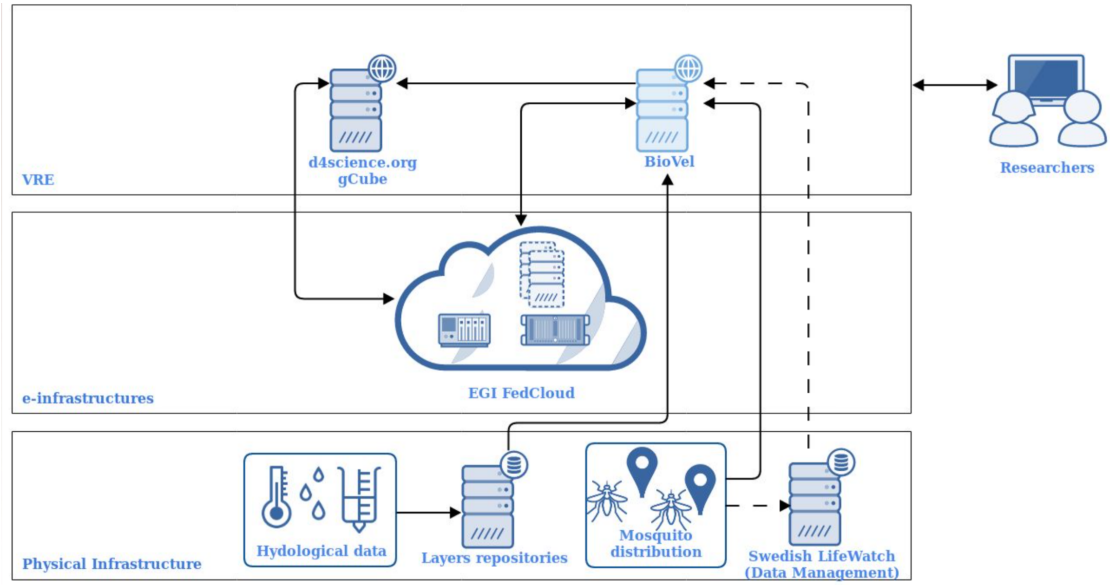


FIGURE 17. USE CASE ARCHITECTURE INCLUDES THREE LAYERS: 1) PHYSICAL INFRASTRUCTURE, 2) E-INFRASTRUCTURE, AND 3) VRE

At the middle layer is the EGI e-infrastructure, which provides Cloud computation and storage resources supporting data-intensive workflow executions.

At the top layer is the D4Science VRE and the Biodiversity Virtual e-Laboratory (BioVel) portal, that provide high-level user interfaces. BioVel[49] is a software environment that assists scientists in collecting, organising, and sharing data processing and analysis tasks in biodiversity and

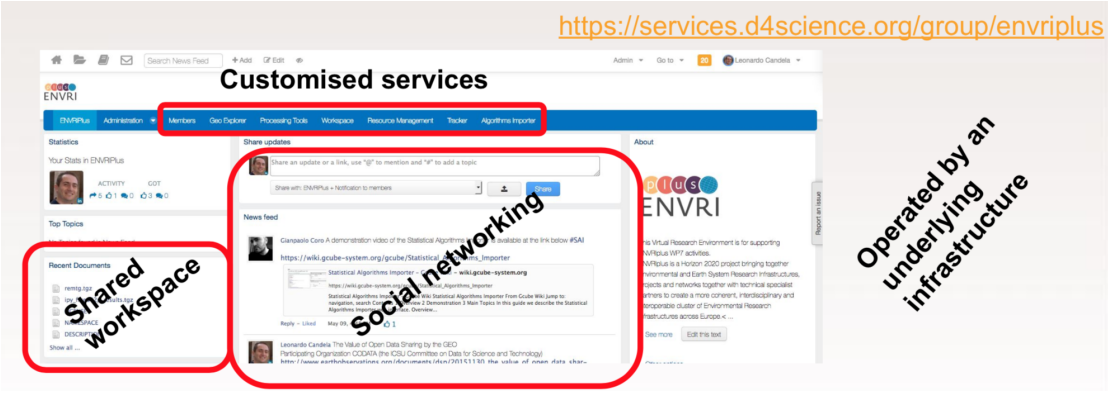
[49] BioVel: https://www.biovel.eu/

ecological research. The service components of the platform include a Biodiversity Catalogue (a library with well annotated data and analysis services), the data processing environments (such as RStudio for creating R programs), a workbench (for assembling data access and analysis pipelines), the myExperiment workflow library (that stores existing workflows), and the BioVel Portal (that allows researchers and collaborators to execute and share workflows).

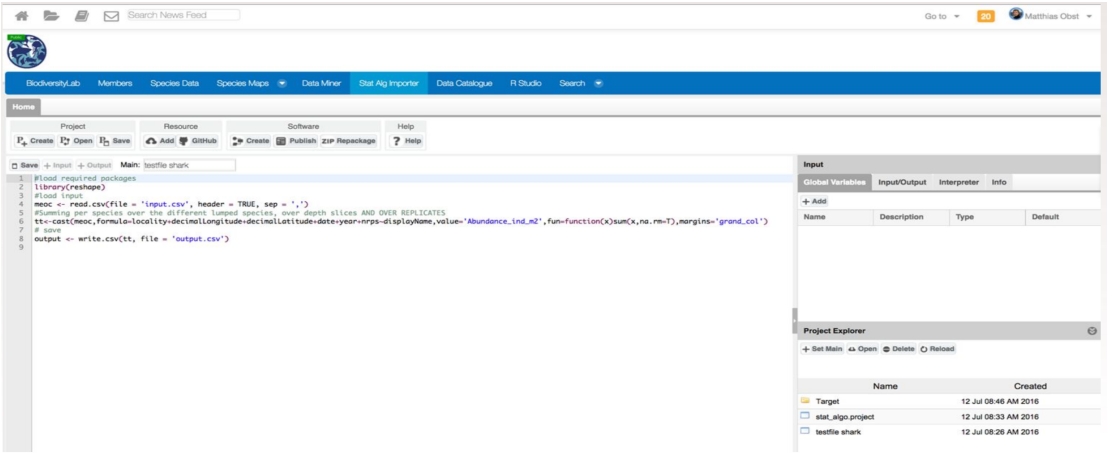
The existing BioVel platform can generate environmental values from species occurrences, however, it only provides standard analysis algorithms. Integrating the D4Science and gCube - based VRE can enrich the functionality of the LifeWatch ICT to allow dynamic modeling.

User Interface

The D4Science/gCube-based VRE for mosquito disease study has been set up with the support from T7.1. The interfaces are shown in Figure 18. It provides a programming environment (shown in Figure 18, b), and it allows biodiversity researchers to develop and compile own/customised analysis algorithms using R, CLI etc. A researcher can decide to share his/her data, algorithms, or workflows by publishing it in the group area (shown in Figure 18, a) that enables social communications via messages, comments, etc.



a VRE AREA FOR SHARING DATA, ALGORITHMS, AND WORKFLOWS



b. VRE AREA FOR DEVELOPING AN ANALYSIS ALGORITHM

FIGURE 18. VRE INTERFACES FOR MOSQUITO DISEASE STUDY

Advantages

Using the VRE, there is no more need for manual sharing of data and algorithms. Information is always synchronized, and data and algorithms are joined in a single place. Users can enjoy an easy and user friendly access interface. The D4Science/gCube-based VRE has an interface to EGI Cloud/HTC resources. If needed, it can outsource the computation on the large-scale e-Infrastructure that can handle computation in parallel and store and share large volumes of data.

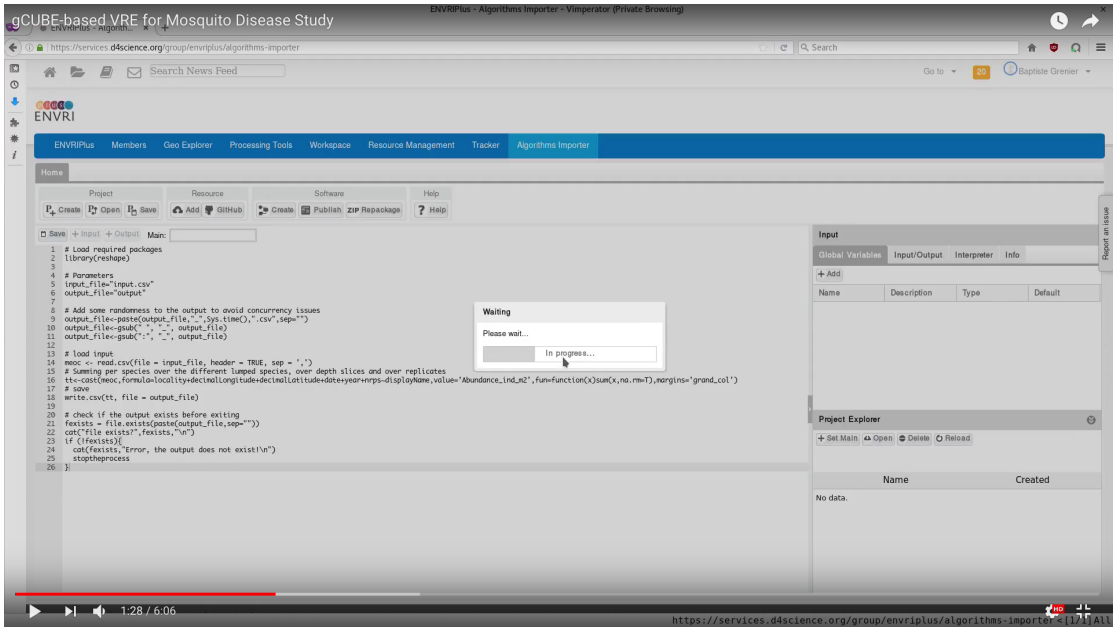
The integration service can bring added value to the Lifewatch community. It makes it possible for individual researchers to repeat and reuse algorithms at will, run trend analysis, and add new parameters and custom data. The VRE provides provenance registration that improves reproducibility. The VRE also allows retention of computation results in the user's workspace. This makes it possible to edit and adapt algorithms.

The integration service also brings added value to ENVRIplus community. Enabling individual researchers to share data and/or algorithms is common to many ENVRIplus RIs where currently data is processed using standard models. Researchers want to use different analysis models and they need a VRE to work together.

This pilot investigation tested and validated WP7 technology. The demo illustrates the integration solutions of linking gCube VRE to LifeWatch RI and to the EGI e-Infrastructure. There are also some lessons learned from the pilot activities: The D4Science/gCube VRE is easy for simple algorithms . It needs integration efforts for complicated algorithms , that requests domain researchers to have technical skills to work with different technology.

Link to the Demonstrator

This demonstrator illustrates a proof of concept of the proposed architecture that uses D4Science to setup a community-centric VRE, allowing researchers to share a simple algorithm with some data and allowing running computations on EGI FedCloud e-Infrastructure. The data was gathered and produced by the involved RIs and the dynamic computation was executed on EGI Federation's resources.



Link: https://youtu.be/IBJkSys5tVo

Contributors

- Baptiste Grenier, EGI, baptiste.grenier@egi.eu
- Matthias Obst, Swedish Lifewatch, matthias.obst@marine.gu.se
- Leonardo Candela, CNR-ISTI, leonardo.candela@isti.cnr.it
- Gianpaolo Coro, CNR-ISTI, gianpaolo.coro@isti.cnr.it