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

Please provide your feedback on this Science Demonstrator using the questionnaire at https://survey2.icos-cp.eu/ENVRIplus-evaluator!

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 1. 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).
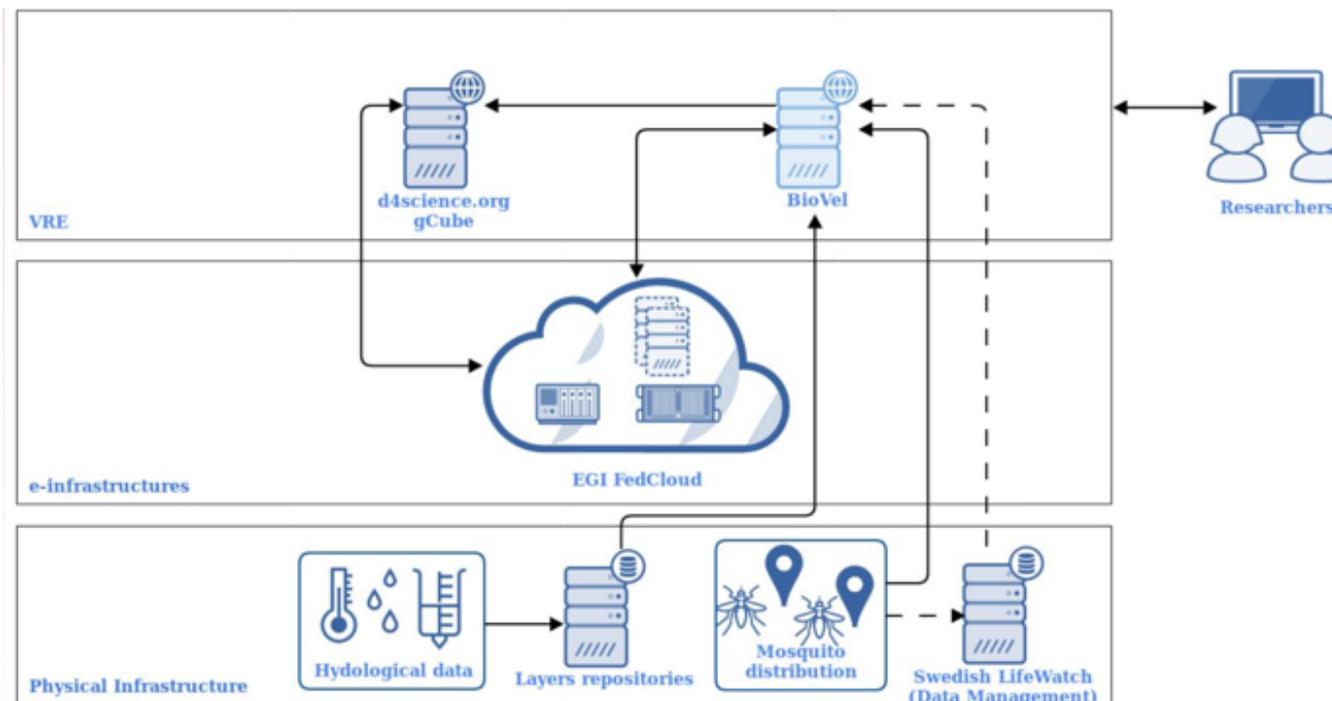

Figure 1. Use Case Architecture includes three layers: 1) Physical Infrastructure, 2) e-Infrastructure, and 3) VRE

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 is a software environment that assists scientists in collecting, organising, and sharing data processing and analysis tasks in biodiversity and 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 2. It provides a programming environment (shown in Figure 2, 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 2, a) that enables social communications via messages, comments, etc.
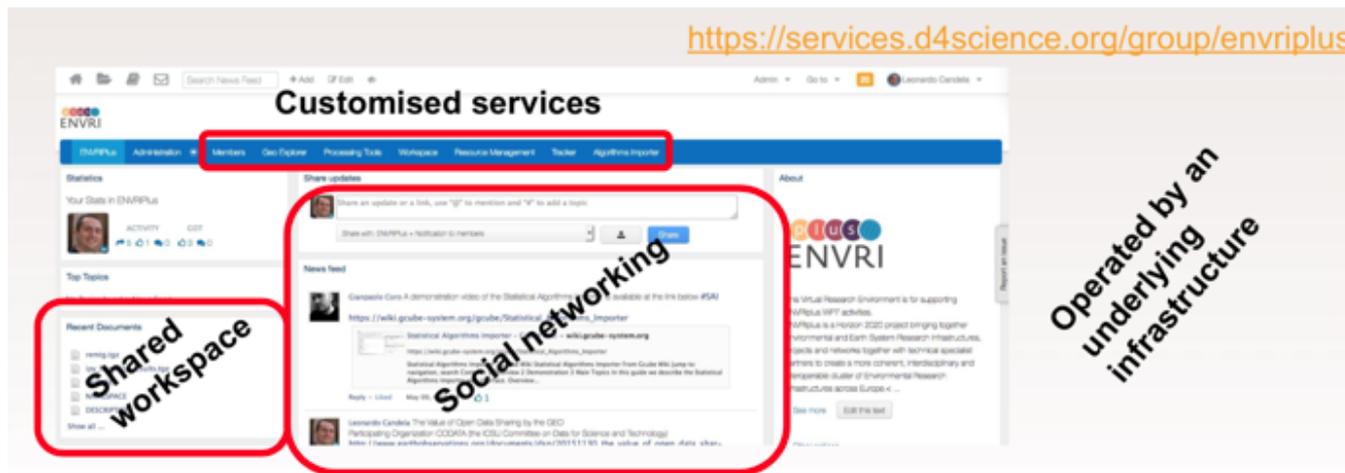
a. VRE area for sharing data, algorithms, and workflows
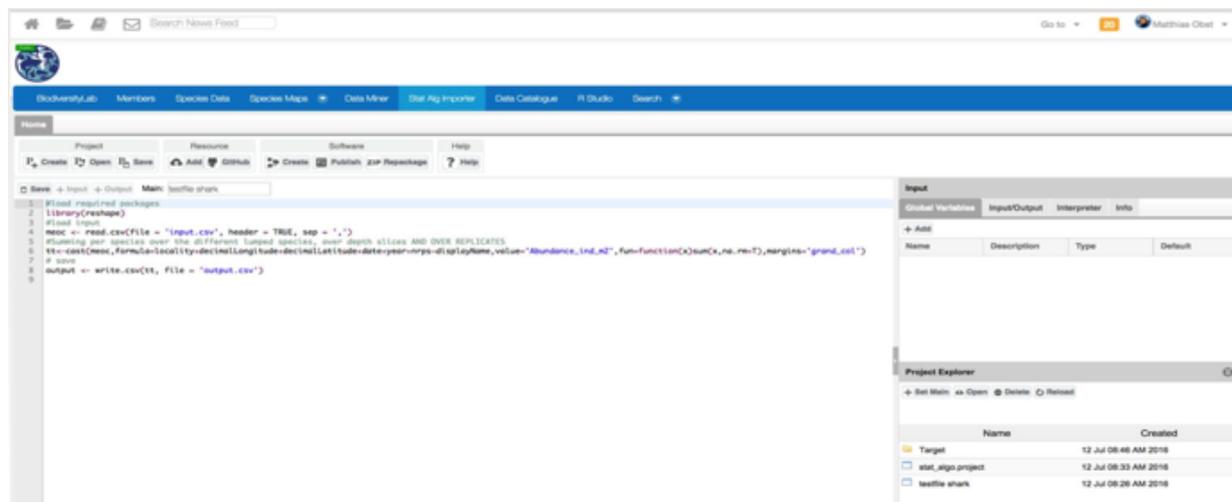
1. **b. VRE area for developing an analysis algorithm**

Figure 2. 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.

Youtube video is at https://youtu.be/XtcCxkFX98I

Contributors

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