EOSC-hub

EOSC Early Adopter Programme

Final Report

Abstract

The EOSC Early Adopter Programme (EAP) is a programme launched by EOSC-hub for research communities interested in exploring the latest state-of-art technologies and services offered by the European Open Science Cloud (EOSC).

13 pilots were selected in two calls spanning several scientific disciplines: agriculture, marine, material sciences, disaster mitigation, health science, biology, astronomy, light pollution etc. The pilots exploited services and resources from EOSC-hub and its partners and completed 25 integrations of 9 different EOSC services.

The majority of these pilot activities will continue after the end of EOSC-hub thanks to agreements with EGI and EUDAT or supported by INFRAEOSC-07 follow-up projects (EGI-ACE, DICE, C-SCALE) tangibly proving how an operational EOSC can boost the research in Europe on adopting new paradigms to deal with the increasing complexity of the science.

EOSC-HUB RECEIVES FUNDING FROM THE EUROPEAN UNION'S HORIZON 2020 RESEARCH AND INNOVATION PROGRAMME UNDER GRANT AGREEMENT NO. 777536.

COPYRIGHT NOTICE

The EOSC-hub project is co-funded by the European Union Horizon 2020 programme under grant number 777536.

TERMINOLOGY

https://wiki.eosc-hub.eu/display/EOSC/EOSC-hub+Glossary

1 Contents

1 Introduction 7

2 Towards an e-infrastructure for plant phenotyping 9

2.1 About the pilot - initial ambition 9

2.2 Progress and key results 9

2.3 Service integrations 10

2.4 Lessons learnt 11

2.5 Impact 11

2.6 Future plans and sustainability aspects 11

3 STARS4ALL 12

3.1 About the pilot - initial ambition 12

3.2 Progress and key results 13

3.3 Service integrations 14

3.4 Lessons learnt 14

3.5 Impact 14

3.6 Future plans and sustainability aspects 15

4 Big Data Analytics for agricultural monitoring using Copernicus Sentinels and EU open data sets 16

4.1 About the pilot - initial ambition 16

4.2 Progress and key results 16

4.3 Service integrations 18

4.4 Lessons learnt 18

4.5 Impact 19

4.6 Future plans and sustainability aspects 19

5 Transitioning EMSO ERIC Data Management Platform to production 20

5.1 About the pilot - initial ambition 20

5.2 Progress and key results 21

5.3 Service integrations 21

5.4 Lessons learnt 22

5.5 Impact 22

5.6 Future plans and sustainability aspects 23

6 Mapping the sensitivity of mitigation scenarios to societal choices 24

6.1 About the pilot - initial ambition 24

6.2 Progress and key results ... 24
6.3 Service integrations ... 25
6.4 Lessons learnt ... 26
6.5 Impact ... 26
6.6 Future plans and sustainability aspects ... 26
7 Open AiiDAlab platform for cloud computing in Materials Science ... 27
7.1 About the pilot - initial ambition ... 27
7.2 Progress and key results ... 27
7.3 Service integrations ... 27
7.4 Lessons learnt ... 28
7.5 Impact ... 28
7.6 Future plans and sustainability aspects ... 28
8 VESPA-Cloud ... 30
8.1 About the pilot - initial ambition ... 30
8.2 Progress and key results ... 30
8.3 Service integrations ... 31
8.4 Lessons learnt ... 32
8.5 Impact ... 32
8.6 Future plans and sustainability aspects ... 32
9 OpenBioMaps data management service for biological sciences and biodiversity conservation 33
9.1 About the pilot - initial ambition ... 33
9.2 Progress and key results ... 33
9.3 Service integrations ... 33
9.4 Lessons learnt ... 34
9.5 Impact ... 34
9.6 Future plans and sustainability aspects ... 35
10 Towards a Global Federated Framework For Open Science Cloud: Three Use Cases ... 36
10.1 About the pilot - initial ambition ... 36
10.2 Progress and key results ... 36
10.3 Service integrations ... 37
10.4 Lessons learnt ... 38
10.5 Impact ... 38

10.6 Future plans and sustainability aspects 38

11 Supporting FAIR data discoverability in clinical research: providing a global metadata repository (MDR) of clinical study objects 39

11.1 About the pilot - initial ambition 39

11.2 Progress and key results 41

11.3 Service integrations 42

11.4 Lessons learnt 43

11.5 Impact 44

11.6 Future plans and sustainability aspects 45

12 EOSC DevOps framework and virtual infrastructure for ENVRI-FAIR common FAIR data services 46

12.1 About the pilot - initial ambition 46

12.2 Progress and key results 46

12.3 Service integrations 47

12.4 Lessons learnt 47

12.5 Impact 48

12.6 Future plans and sustainability aspects 48

13 Integration of toxicology and risk assessment services into the EOSC marketplace 49

13.1 About the pilot - initial ambition 49

13.2 Progress and key results 50

13.3 Service integrations 51

13.4 Lessons learnt 51

13.5 Impact 52

13.6 Future plans and sustainability aspects 52

14 AGINFRA+: Virtual Research Environments to Support Agriculture and Food Research Communities 53

14.1 About the pilot - initial ambition 53

14.2 Progress and key results 53

14.3 Service integrations 54

14.4 Lessons learnt 54

14.5 Impact 54

14.6 Future plans and sustainability aspects 55

Executive summary

The EOSC Early Adopter Programme (EAP)[1] is a programme launched by EOSC-hub for research communities interested in exploring the latest state-of-art technologies and services offered by the European Open Science Cloud (EOSC).

The programme was thought for researchers and research communities working on complex research projects in need of services based on multiple, different technologies, that wanted to benefit from access to state-of-the-art technologies, research infrastructures and services that were not available in their current research environment. Selected research communities were able to scale up their in-house infrastructure and to access and combine a richer set of services and resources, through the EOSC Portal & Marketplace[2], for data discovery and reuse, data processing and analysis, data management, curation and preservation, and access, deposition and sharing, federated authentication and authorization.

The services and resources in scope for the programme were those provided by the EOSC-hub project and the Early Adopter Programme partners, namely OCRE, OpenAIRE and GÉANT. Furthermore EOSC-hub, with its network of technical experts, trained and supported researchers to enable active usage of the EOSC fostering a culture of co-operation between researchers and EOSC providers.

13 pilots were selected in two calls spanning several scientific disciplines: agriculture, marine, material sciences, disaster mitigation, health science, biology, astronomy, light pollution etc. The principal investigators of the pilots were supported by shepherds from the EOSC-hub technical support team that closely guided the researchers on accessing and using the services. At the end, the pilots successfully exploited several services and resources completing 25 integrations of 9 different services. These integrations enabled the setup of new services for research in Europe or the enrichment of existing services. In such way, the programme increased the confidence of the researchers in the capability and capacity that will be provided via the EOSC and enabled several long-term collaborations between user communities and EOSC service providers.

Key examples of the success of the programme are the EMSO pilot, that allowed the EMSO ERIC to make operational its data management platform[3] thanks to the deployment in the EGI Federated Cloud and the integration with the EGI AAI Check-in service, and the Big Data Analytics for agricultural monitoring pilot, that demonstrated how large datasets like the Copernicus EO data can be properly managed and exploited federating compute and storage resources in EOSC.

As a result, the majority of these pilot activities will continue after the end of EOSC-hub thanks to agreements with EGI and EUDAT or supported by INFRAEOSC-07 follow-up projects (EGI-ACE, DICE, C-SCALE) tangibly proving how an operational EOSC can boost the research in Europe on adopting new paradigms to deal with the increasing complexity of the science.

[1] https://eosc-portal.eu/news/new-booklet-describing-eosc-early-adopter-programme
[2] https://eosc-portal.eu/
[3] https://marketplace.eosc-portal.eu/services/emso-eric-data-portal

1 Introduction

This document is the final report of the EOSC Early Adopter Programme (EAP) that was launched by EOSC-hub to explore the latest state-of-art technologies and services offered by the European Open Science Cloud (EOSC).

The EOSC EAP selected 13 pilots in two calls and supported them for a period of around 1 year.

The following table lists and shortly describes the 13 EAP pilots. More information is available in the next sections where for each pilot we presented initial ambition, progresses and key results, achieved integrations, lesson learnt, impact, future plans and sustainability aspects.

Table 1. Pilots of the EOSC Early Adopter Programme.

Pilot	Institution	Scientific area
Towards an e-infrastructure for plant phenotyping	INRA (France)	Plants and agriculture
Mapping the sensitivity of mitigation scenarios to societal choices	IIASA (Austria)	Earth and related Environmental sciences, Economics and Business
STARS4ALL	STARS4ALL Foundation	Light Pollution
Transitioning EMSO ERIC Data Management Platform to production	EMSO ERIC	Earth Sciences
Big Data Analytics for agricultural monitoring using Copernicus Sentinels and EU open data sets	European Commission, Joint Research Centre	Earth and Environmental Sciences, Agriculture
Supporting FAIR data discoverability in clinical research: providing a global metadata repository (MDR) of clinical study object	ECRIN (France)	Health Sciences
Open AiiDA lab platform for cloud computing in Materials Science	EPFL (Switzerland)	Physical sciences, Chemical sciences, Materials engineering

VESPA-Cloud	OBSPM (France)	Astronomy
OpenBioMaps data management service for biological sciences and biodiversity conservation	UNIDEB (Hungary)	Biology, conservation biology, ecology, biodiversity
AGINFRA+: Virtual Research Environments to Support Agriculture and Food Research Communities	CNR (Italy)	Agricultural sciences
EOSC DevOps framework and virtual infrastructure for ENVRI-FAIR common FAIR data services	ENVRI-FAIR Cluster project	Earth and related Environmental sciences
Integration of toxicology and risk assessment services into the EOSC marketplace	Edelweiss Connect GmbH (Switzerland)	Bioinformatics
Towards a Global Federated Framework For Open Science Cloud: Three Use Cases	AASCTC (Saudi Arabia) & CNIC-CNAS (China)	Several areas

2 Towards an e-infrastructure for plant phenotyping

Principal investigator: Vincent Nègre (INRAE)

Shepherd: Nicolas Cazenave (CINES)

2.1 About the pilot - initial ambition

In recent years, technological progress has been made in plant phenomics (major improvements concerning imaging and sensor technologies). High-throughput plant phenotyping platforms now produce massive datasets involving millions of plant images concerning hundreds of different genotypes at different phenological stages in both field and controlled environments. Networks of sensors also measure environmental conditions in real time. The ongoing robotization of experimental processes foreshadows an explosion in the volume and complexity of the data produced by the different research facilities. There is a need for an integrated and federated solution for data management and data processing.

The open-source Phenotyping Hybrid Information System PHIS (Neveu et al. 2019 New Phytologist, 221: 588–601) has been developed to organize these data and make them accessible and reusable to a larger scientific community.

Three use cases have been proposed to explore which EOSC-hub services are the most appropriate to support a European plant phenotyping e-infrastructure.

- Use case 1: the PHIS information system and the Galaxy environment will be deployed on **EGI virtual machines.** The storage layer is based on the existing **FranceGrilles iRODS** infrastructure. An authentication layer based on the **EGI Check-in service** and a computing layer provided with the **EGI Notebooks service** will be added.
- Use case 2: **the storage layer is based on the B2SAFE service** supported by the EUDAT infrastructure.
- Use case 3: the **storage layer is based on the Data Hub service** supported by the EGI infrastructure.

2.2 Progress and key results

The early adopter program allowed us to deploy our first pilot based on several services provided by EOSC.

Our information system is hosted on a virtual machine hosted by CESNET-MCC provider. The EGI Check-in service has been integrated as the default authentication system. This service is very useful because it allows users to connect with their institutional identifiers in a transparent way.

We have also connected our information system to the online storage service provided by IN2P3-IRES and the FranceGrilles e-infrastructure. This service is based on the distributed storage system iRODS. We have configured the iRODS system to do automatic replication on different data centers to secure the data. Users can transfer their files transparently via our web services that communicate with irods through the commands.

We are currently working on a second pilot to integrate the online storage system based on the EGI Data Hub and the Onedata distributed storage system. A Onedata instance has been installed on the CESNET-MCC and CYFRONET-CLOUD infrastructures. An automatic data replication has been set up between those 2 sites. The integration with our information system should be possible thanks to the interfaces provided by OneData (a REST API and a Cloud Data Management Interface).

In order to increase the visibility of our services we want to onboard them in the EOSC Portal. We are not yet ready for that because we have to finalize the integration of Onedata service; and we have to integrate public data. We will apply to the EGI-ACE program in order to continue in this direction.

2.3 Service integrations

Service	Provider	Description of the integrations	Allocated ICT resources (cloud, storage, etc.)
EGI Cloud Compute	EGI	Deployment of our information system	CESNET-MCC: 24 vCPUs; 192GB RAM CYFRONET-CLOUD: 8 vCPUs; 32 GB RAM; 80 GB local disk IN2P3-IRES: 24 vCPUs, 64GB RAM; 40 GB disk
Online storage	EGI	Onedata instance & iRODS	CESNET-MCC: 10Tb storage capacity CYFRONET-CLOUD 10Tb storage capacity IN2P3-IRES: 10Tb storage capacity
EGI Check-in	EGI	Federated identity management system based on EGI Check-in	

2.4 Lessons learnt

The Early Adopter Program allowed us (i) to understand better how the EOSC infrastructure works, (ii) to identify useful services for our community. The exchanges and feedback from other projects involved in the EAP program were also very enriching.

This first experience with the EOSC was very positive. We are convinced that the EOSC portal meets the needs of the plant phenotyping community, especially at the European level where the landscape is fragmented. The EOSC provides a solution for better integration and interoperability between partners.

2.5 Impact

The impact has been positive on several levels. The EAP program allowed us to:

- access to a robust infrastructure;
- facilitate access to our services to the European partners;
- benefit from innovative services;
- increase our visibility by publishing our services into the EOSC portal.

2.6 Future plans and sustainability aspects

- We would like to maintain the access to the services we have tested in the project. That would give us the opportunity to finalize the deployment of our second pilot and publish a demo version with public data on the EOSC portal. We have already initiated the discussion to the service providers . As the CYFRONET provider would like to continue to support your use case through the EGI-ACE program we will submit our application in the next call.
- We would also like to continue testing new services offered by EOSC. We are interested in services to process data (in particular to GPU resources). We are also looking for services that could be proposed for the semantic web. As we need some time to improve our proposal in this direction, we will not apply for the first call (deadline is 15/04) but for the next one (in 2 months).

3 STARS4ALL

Principal investigator: Esteban Gonzalez (STARS4ALL Foundation)

Shepherd: Daan Broeder (Meertens Institute)

3.1 About the pilot - initial ambition

STARS4ALL is a community concerned with the light-pollution problem. The community derives its name from the STARS4ALL project funded by the European Union H2020 Programme (688135) that was created to create awareness among citizens about the light pollution problem. For this purpose, it deploys a platform to give support to some light pollution initiatives. These initiatives include a photometer network (http://tess.stars4all.eu) to continuously monitor the light pollution, using photometers to measure the sky brightness. In this context, it deploys a platform (http://tess.dashboards.stars4all.eu) to show the measurements in some dashboards. Besides the photometer network, STARS4ALL gives support to citizen science initiatives like Cities At Night. All data is published openly in the STARS4ALL Zenodo community (https://zenodo.org/communities/stars4all). The project was completed in 2018 but STARS4ALL continues the work through the STARS4ALL foundation created for this purpose.

The STARS4ALL Pilot goals are to make the current infrastructure more robust and provide a higher capacity, through the EGI Cloud Compute services, and improve discoverability and data access of the project and its data via the general EOSC-hub services.

Using:

- B2SHARE for storing and providing access to STARS4ALL primary data and derived products, using an appropriate STARS4ALL community metadata schema;
- B2FIND for improving discoverability of STARS4ALL data, relying on the metadata harvesting by B2FIND from B2SHARE;
- GEOSS portal, as a specialised discovery platform for Earth Observation Data as an additional discovery option for STARS4ALL data. The harvesting of STARS4ALL metadata was planned through the B2SHARE OAI-PMH provider service.

Improve data management practices by:

- facilitating depositing of also secondary and tertiary data (analysis data and publications);
- provide Resource Objects/Bundles for related primary, secondary and tertiary data products;
- introduction of PIDs for the STARS4ALL sensor equipment (photometers).

Improving accessibility and usability of the STARS4ALL data by users:

- actionable links to data objects will be provided via B2FIND and B2SHARE

New data analysis options will be provided by Jupyter Notebooks (using the EGI Notebook service) for analysing the observation data directly from B2SHARE and ZENODO via their respective APIs.

Also discussion and investigations of the usability of additional data management services as provided in the EOSC portal was part of Pilot activities.

3.2 Progress and key results

Results and progress of the STARS4ALL pilot are the following.

Increasing the robustness of the infrastructure was achieved by duplicating the essential infrastructure services first via the EGI Cloud Compute, but later through facilities provided by a local university collaboration. However the use of EGI Cloud based load-balancer was continued.

The use of B2SHARE for storing the STARS4ALL generated data was successful, however we were not able to use production B2SHARE service in combination with the latest STARS4ALL community metadata schema, since the B2SHARE support postponed the community metadata schema modifications until the B2SHARE core metadata schema could be updated (expected March 2021). All work was therefore performed using the B2SHARE test-instance (https://eudat-b2share-test.csc.fi/communities/STARS4ALL) with a limited community metadata schema. Which was sufficient for testing and demonstrating the advantages of using B2SHARE including its flexible support for community metadata schema and direct access from Jupyter Notebooks to the B2SHARED stored data. However, we noticed the limited interoperability of such community metadata with the other EOSC-hub services e.g. the B2FIND metadata harvesting and the GEOSS portal. This was mentioned as a reason for the new B2SHARE core metadata schema, but cannot be tested before the end of this pilot.

Regarding the data management, the use of the new schema in B2SHARE has made possible to publish our datasets grouped by sensor. That is to say, our users can access all datasets generated by a specific sensor. Furthermore, through the search engine of B2SHARE, we can access all the results generated by sensors located in a country, or sensors with a specific filter. It should also be pointed out that all datasets have been published using the API provided by B2SHARE. This is key to automatize the project because our photometer network is generating data continuously.

The pilot also investigated which existing PID services would be suitable to be used to provide (resolvable) identifiers for the STARS4ALL photometer sensors. Considered were DOIs, the EPIC Handle service (B2HANDLE), and operating an own STARS4ALL Handle service. The requirements for such a service included the wish to own the Handle prefix (ie. independence of the service provider long-term), being able to control the Handle resolving and affordable costs. Our investigation pointed to using B2HANDLE, and testing the B2HANDLE API proved successful. However, the annual costs proved prohibitive for a small project as STARS4ALL. Also, the new (experimental) B2Inst service was also investigated, however no possibilities for control of the Handle resolving seemed to exist. Operating an own STARS4ALL Handle service is still under consideration.

In the context of investigating other potential useful services, we experimented with Virtual Collections as provided by the CLARIN Virtual Collection Registry. This is an EOSC-hub thematic service operated by the CLARIN ERIC and allows the creation and publication of collections of heterogeneous and distributed data. One of the particularities of our network is that it is composed by photometers operated by other projects. We established its usefulness for grouping photometers in collections for each project.

Finally, Jupyter notebooks have been created to exploit the data published in B2SHARE, through the EGI Notebooks service. These networks can be used for scientific purposes (to calculate light sources close to a sensor) and educational purposes (to visualize the duration of the daylight during the year depending on the latitude).

3.3 Service integrations

Service	Provider	Description of the integrations	Allocated ICT resources (cloud, storage, etc.)
B2SHARE	EUDAT CDI	Publications of the datasets generated in our photometer network	Archiving
EGI Cloud Compute	EGI	Deployment of a load balancer to increase the availability of our platform	CESGA: 4 vCPUs, 24GB RAM; 1 public IP
EGI Notebooks	EGI	Deployment of scientific and educational notebooks	CESGA: Number of users: 4 Cores per user: 2 Memory per user (GB): 4GB Storage per user (GB): 40GB
Virtual Collections	CLARIN	Creation of collections of photometers	

3.4 Lessons learnt

The need to provide metadata interoperability between the different EOSC services, at least on the level of a core-set that includes basic discovery via geo-spatial parameters.

The need for a free PID service that fits the limited financial means of citizen-science projects as STARS4ALL.

3.5 Impact

- Increased robustness and capacity of the STARS4ALL infrastructure;

- increased discoverability of STARS4ALL data, although the final production workflow still need to be realised;
- awareness of other potential helpful data management practices and services e.g. CLARIN Virtual Collection Registry.

3.6 Future plans and sustainability aspects

The EGI Notebook service to setup our Jupyter Notebook was successfully adopted, a proper SLA was offered, and it will be considered also for future use.

Although we did not receive a specific SLA offer for continued B2SHARE support beyond the end of the EOSC-hub project, the B2SHARE support team has assured us that the research communities will receive continued support also after the end of the EOSC-hub project. Our investigation pointed to using B2HANDLE and testing the B2HANDLE API proved successful. However, the annual costs proved to be prohibitive for a small project as STARS4ALL.

4 Big Data Analytics for agricultural monitoring using Copernicus Sentinels and EU open data sets

Principal investigator: Guido Lemoine (EC JRC)

Shepherd: Enol Fernandez (EGI.eu)

4.1 About the pilot - initial ambition

Earth Observation (EO) is a relatively new domain on the European Open Science Cloud. EO data access has long suffered from restrictive licenses and opaque and proprietary distribution systems, which has, by and large, hindered wide uptake in science, in particular beyond the traditional remote sensing and geospatial analysis disciplines. Massive new EO data streams which are distributed under a full, free and open license include those from the European Copernicus program's Sentinel sensors since 2014 and US Landsat since 2010. Currently multiple Petabytes of high resolution Sentinel-1 (SAR) and -2 (optical) sensor data are available for thematic research and monitoring applications in maritime and land science disciplines.

Still, even with open licenses, EO data access remains complex and combining such data with geospatial reference data for targeted analysis is hard for novice users. Extensive knowledge of sensor-specific data organization, map projections and formats is often required. Some data, for instance Sentinel-1, requires complex processing to create "analysis ready" data sets. The old paradigm in EO data analysis was that 80% of a researcher's time was spent on data pre-processing and preparation, and 20% on analysis. This radically changed with the introduction of cloud infrastructure that hosts complete sensor data collections closely coupled with massive parallel processing capacity. By abstracting data access and integrating ever more sophisticated analysis geospatial analysis routines, science users are able to compose their analysis in scripts that can be executed interactively or in batch. This allows the paradigm to be inverted, i.e. 80% of research time can potentially be spent on programming and testing the analytical logic that underlies scalable and reproducible science methods.

This pilot builds on expertise from developments in Google Earth Engine and Copernicus Data and Information Systems (DIAS) used in agricultural monitoring for EU Common Agricultural Project requirements. A key objective is to demonstrate that functionalities developed in this context can be integrated in a federated cloud framework that addresses needs of scientific data users as well, in particular those that require applications of novel machine learning to very large geographical feature sets and deep time stacks.

4.2 Progress and key results

The EAP pilot has allowed us to quickly build out the initial code base and structure it in a set of (python) modular components that address backend and frontend functionalities. The scale up to use complete national coverage was greatly supported by federating the CloudFerro (CREODIAS) DIAS core processing facilities on which we applied parallelization with docker swarm with dedicated PostgreSQL database servers and server components on CESNET infrastructure. EODC resources were effectively used to generate application ready data for Sentinel-1 using alternative processing recipes, e.g. for partial polarization and radiometric terrain corrected outputs. Front-end functionalities address the needs of data analytics and visualization and reporting needs in both operational and scientific use contexts. They rely on back-end server components like JupyterHub and RESTFul services to give access to the time series databases and large S3 object storage on the DIAS instance. An overview of the system setup is given in Figure 1. We have both tested a dockerized JupyterHub set up on CESNET and the EGI JupyterHub. For RESTful services we deploy a dockerized Flask container on CESNET. Towards the end of 2020, the code base was open sourced and placed on a github.com/ec-jrc/cbm. The code comes with ample documentation.

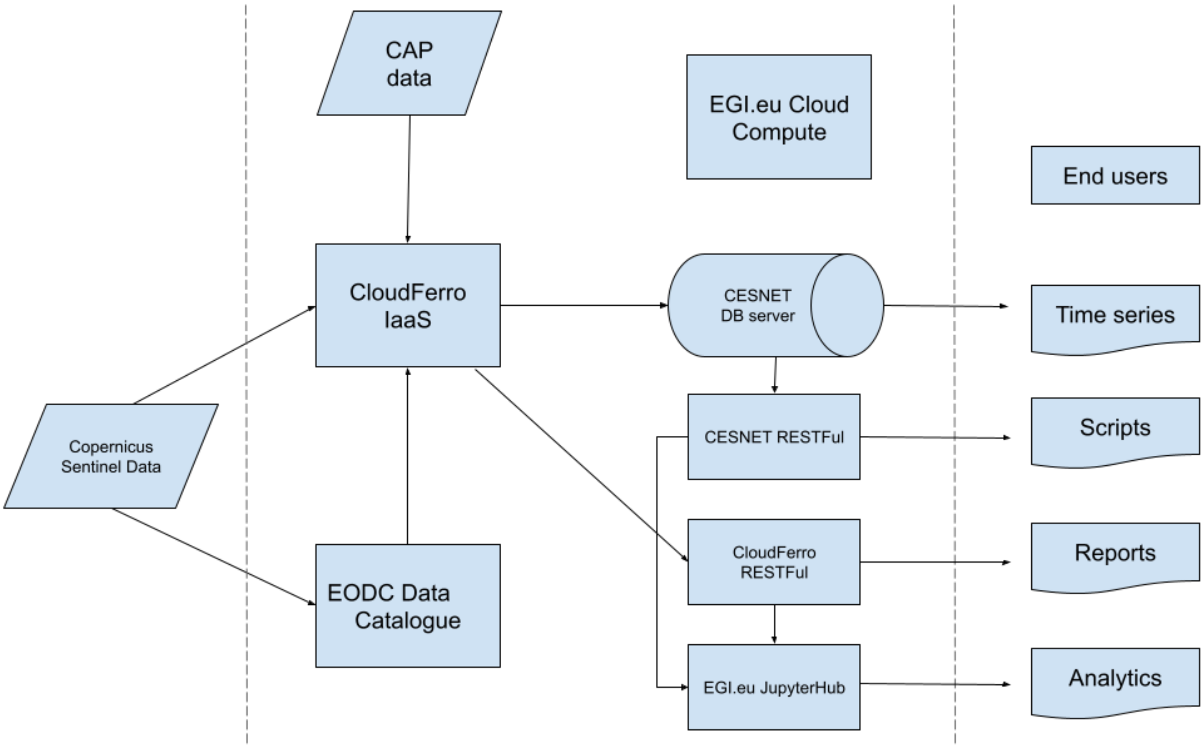


Figure 1. A schematic overview of the EOSC portal services used in this Early Adaptor Project. CloudFerro and EODC are part of the EOSC-hub EO pillar Thematic service.

We have achieved TRL 6-7 for several operational tasks in the CAP monitoring context. Application Ready Data processing for Sentinel-1 is now offered as a TRL-9 Processing-as-a-Service (PaaS) module on CREODIAS at unit pricing (was TRL 7 at start of the project). Some functions (e.g. generic image subsetting for “calendar view” applications) are at TRL-7 and were discussed with CloudFerro for implementation as long term sustainable DIAS functions. Another function for a PaaS-based ARD image stack extraction of territorial agricultural parcel sets is currently under discussion. The integration of these offerings in the EOSC Portal requires concluding these discussions. Whereas database components used in the project were satisfying the project needs, some additional thought is needed to build this out as a service infrastructure that could serve an arbitrary number of projects of a similar nature. This is mostly linked to database organization and clustering over scalable storage and server architecture, for which limited knowledge existed in the project.

4.3 Service integrations

Service	Provider	Description of the integrations	Allocated ICT resources (cloud, storage, etc.)
CloudFerro Infrastructure	CloudFerro	Integrates access to multiple Petabyte S3 store of Copernicus Sentinel data with cloud compute	24 vCPUs, 32 GB RAM, 100 GB HDD
EGI Cloud Compute	EGI	Server backend setup on performant cloud resources with large data storage (CEPH based HDD + S3)	CESNET: 16 vCPUs, 64 GB RAM, 100 GB HDD 32 vCPUs, 128 GB RAM, 200 GB HDD Database server, 4 vCPUs, 16 GB RAM, 1 TB HDD 20 TB S3 storage
EODC Data Catalogue Service	EODC	Querying and processing of Copernicus Sentinel-1 data and compute solutions for ARD	8 vCPUs, 16 GM RAM, 100 GB HDD

4.4 Lessons learnt

The EAP pilot has boosted our confidence that a European federated Open Science Cloud provides the essential compute needs of the European Earth Observation community, both in the science domain and in the public use domain. The transparent integration of industry strength infrastructure (e.g. DIAS) with advanced scientific compute solutions correctly addresses the need for a coherent solution in a very fragmented European landscape. The potential to create a more level playing field between applied science and operational practices will facilitate much faster uptake of novel ideas and analytical approaches, esp. those emerging from machine learning and data integration from different agricultural stakeholders and addressing needs across the domain, including those of individual farmers.

The initial participation of scientific partners (Wageningen Environmental Research and University Twente) did not materialize in active collaboration. However, awareness amongst users in the CAP monitoring community has grown very significantly, both in terms of DIAS use but also in the understanding of the needs for (public) European processing solutions that best integrate and simplify access to federated scientific and industry compute capacities that serve the broadest possible communities of practice [in agriculture]. The recent introduction of the European Green Deal and related discussion on digitization needs in the European agricultural domain are of great relevance in this respect.

4.5 Impact

The EAP pilot has allowed us to rapidly build out functionality that is relevant for our core user community. The cross-link to the science community has been less successful, but requires additional actions, e.g. by making time series extracts available as open data (work in progress).

The pilot has attracted significant interest inside the JRC and discussion has started on how JRC can be associated with EOSC and euroHPC infrastructure programmes. Inside JRC, a large number of compute intensive applications are developed, e.g. in environmental and economic modeling, in machine learning and AI and Big Data analytics. The Commission's open data and open source policies are paving the way to integrate external European compute infrastructure, which will have a significant impact on JRC ICT strategy.

4.6 Future plans and sustainability aspects

We hope to continue development under new projects, such as the recently started C-SCALE and EGI-ACE H2020 projects. Discussion is ongoing on how resource use may migrate under these projects. This will be combined with pay for use DIAS, esp. for the CAP monitoring users across the EU (using direct Commission funding). We are closely following developments in DG CNECT e-infrastructure (euroHPC, EOSC, European Data Spaces) and how they impact our project activities and more generic compute needs of other JRC user domains.

Discussions with EOSC resource providers on the establishment of new PaaS and generic functionalities for a wider community of infrastructure users (in particular DIAS) is ongoing.

5 Transitioning EMSO ERIC Data Management Platform to production

Principal investigator: Ivan Rodero (EMSO ERIC)

Shepherd: Giuseppe La Rocca (EGI.eu)

5.1 About the pilot - initial ambition

The European Multidisciplinary Seafloor and water column Observatory (EMSO) aims to explore the oceans, to gain a better understanding of phenomena happening within and below them, and to explain the critical role that these phenomena play in the broader Earth systems (http://emso.eu/what-is-emso/).

- Provide deep sea high quality, long term time series.
- Develop technology for sensors, communications, offshore operations.
- Attract scientists, technicians, managers and industries.
- Collaborate with European and International Organization and Institution (specifically in EOOS and GEOSS).
- Promote innovation and knowledge-sharing.
- Conduct outreach and communication.

EMSO's observatories are platforms equipped with multiple sensors to measure, for example, biogeochemical and physical parameters such as ocean temperature, dissolved oxygen concentration, and ocean current speed and direction.

A fundamental information technology component of the EMSO cyber-infrastructure, that allows the integration of data from EMSO regional facilities where the observatories are deployed, is the Data Management Platform (DMP). The DMP has been designed to deliver a flexible and scalable data management platform using open source big data frameworks for long-term, high-resolution, (near)-real-time monitoring data by providing a coordinated approach for data capturing, archiving, management and delivery based on OGC standards. The EMSO DMP has adopted a set of common services, and has been complemented with widely-used tools (e.g., federated ERDDAP deployment). The DMP ingests, consolidates, processes and archives data, integrates the data management architectures of the regionally distributed EMSO nodes and makes data available to the community.

This pilot aims at transitioning the DMP to pre-production and facilitating the way to its full production. The prototype DMP has been deployed at EGI, and its current technology readiness is at level 8, and its transition to level 9 is undergoing. This transition enables data and services to be harmonized to bring accessibility and, when consistent and relevant, be enhanced through enriched metadata. The existing heterogeneous and distributed EMSO data web services are being harmonized and standardized across EMSO nodes and made interoperable with the subdomain according to FAIR principles, which has the potential to foster interoperability between EOSC services. It also impacts current efforts within the ENVRI-FAIR H2020 project as it enables EMSO ERIC to establish an appropriate workflow for taking stewardship of every stage of the data lifecycle and ensure long-term preservation and redundancy as well as additional mechanisms for data, metadata, and data product discovery and delivery based on decentralized approaches and standard/widely used tools.

5.2 Progress and key results

The pilot has deployed computing and online storage resources to support three environments:

1. development/test site that provides an environment for software evolution and testing, including configuration management, continuous integration and functional testing,
2. core site that supports the DMP software stack, including back-end processes and data services exposed to users; and
3. backup core site that represents a mirror of the core site for system resiliency and business continuity, including data and services mirroring and fail-over capabilities.

The deployed architecture is based on robustness and fault tolerance, including redundancy and failover capabilities on computing and storage resources, and scalability and security, including a distributed architecture for data access and analysis. Furthermore, solutions for the efficient movement of large datasets and data preservation brings added value to the deployment of the DMP.

The transition of the DMP to pre-production has been crucial as regional facilities and test sites are heterogeneous and distributed and the pilot has enabled the harmonization of data and services, that are key issues for interoperability and reusability. In addition, the processes supported by the DMP enhances the findability and access to data and products. The operational system provides open-access, accurate, long-term measurements of ocean parameters. This, in turn, has led to increased interoperability of EMSO nodes and the consistent collection of ocean essential variables.

EOSC-hub provided cloud-based resources from two geo-distributed datacenters in Italy (RECAS-BARI) and Spain (CESGA) belonging to the EGI Federation, guaranteed by an SLA. Additional critical services have been integrated, such as support for integrating an Authentication and Authorization Infrastructure based on the EGI Check-in service. Services provided through the pilot have also enabled added-value services, including a virtual research environment with data analytics capabilities. Efforts for onboarding of key services such as the EMSO ERIC data portal in the EOSC Portal is undergoing.

5.3 Service integrations

Service	Provider	Description of the integrations	Allocated ICT resources (cloud, storage, etc.)
EGI Cloud Compute	EGI	The EMSO ERIC Data Services has been installed in production at RECAS-BARI. An additional cloud provider	RECAS-BARI: 300 vCPU cores, 1.2TB of RAM

		(CESGA) has been configured to provide fail-over and back-up capabilities	CESGA: 192 vCPU cores, 512GB of RAM
EGI Online Storage	EGI	The EMSO ERIC Data Services has been installed in production at RECAS-BARI. An additional cloud provider (CESGA) has been configured to provide fail-over and back-up capabilities	RECAS-BARI: 10TB of block storage CESGA: 600GB HDD of block storage
EGI AAI Check-in	EGI	Configured the EMSO-ERIC IdP, set-up a federated identity management system based on EGI Check-In	

5.4 Lessons learnt

The pilot has enabled the key milestone of transitioning the data management platform into a pre-production operational system that has been exposed to the community and is allowing gathering precious feedback for its evolution. The integration of essential services such as authentication and authorization infrastructure has accelerated its transition to full production and guarantees the stability of the system.

The use of engineering best practises and deploying an architecture leveraging EOSC services have been essential for achieving robustness and fault tolerance, including redundancy and failover capabilities on computing and storage resources; and scalability and security, including a distributed architecture for data access and analysis.

The integration of additional EOSC services such as monitoring and accounting capabilities are expected to improve operational processes; however, ready-to-use long-term management policies would be a plus.

5.5 Impact

This pilot impacts the community in different dimensions. It provides a robust and production-ready system to find and access curated quality data and data products. It also provides the community with access to services associated with the DMP ecosystem such as analytics and added value services.

The outcomes of the pilot include open access to meaningful, quality and integrated data and data products from EMSO regional facilities. It impacts different stakeholders beyond EMSO ERIC, including researchers, educators, and policy-makers across the globe. The pilot resources also enable EMSO ERIC to provide its community added value services such as analytics, dashboards, and data portals. This data and products are essential for scientific communities in a broad range of domains, including geosciences, biogeochemistry, marine ecology, and physical oceanography.

Open access data and its integration with EOSC impact researchers, educators and the general public from European communities and beyond. During the course of the pilot, EMSO ERIC services operated using EGI resources such as the data portal received thousands of visits from more than one thousand distinct users from 85 countries. The countries with a larger number of visits include China, Italy, Spain, France, Greece, Portugal, United Kingdom, United States, Japan, and Germany.

5.6 Future plans and sustainability aspects

- The EMSO-ERIC Data Service will be registered in the EOSC Portal.
- The Early Adopter will be further supported as a Data Space provider in the context of the EGI-ACE project.
- SLA/OLAs will be extended till 06/2023, supporting the transition of EMSO ERIC data services to full production

6 Mapping the sensitivity of mitigation scenarios to societal choices

Principal investigator: Bas Van Ruijven (IIASA)

Shepherd: Alessandro Costantini (INFN)

6.1 About the pilot - initial ambition

This project aims to perform modeling studies to explore how future energy systems can evolve and to quantify the tradeoffs, co-benefits, and interlinkages between different aspects of the global energy systems in the context of international climate change policy and sustainable development.

Such analyses utilize so-called Integrated Assessment Models (IAMs), which are models of the energy, environment, and economic systems in order to quantify key variables of interest in these scenarios such as emissions pathways consistent with international climate policy goals, tradeoffs of climate mitigation with land use and food security, among others.

This project will provide a proof-of-principle platform aimed at performing large scale (10-15k model runs) analyses.

The IAM **MESSAGEix-GLOBIOM** (considered by the applicants at TRL9) will run sequentially on the selected resources where each job is independent from the other in a parametric fashion. The Model will run in the resources (Virtual Machines) provided by EOSC resource providers. The parametrized simulations will be submitted by making use of a batch system manually deployed by the applicants.

An exploratory activity has been performed by applicants for running the full software stack in a containerized environment (using docker) on larger compute systems (e.g., HTC). Even if this activity is likely TRL5, it is envisioned as a key software infrastructure product to promote to TRL9 during this project. Starting from such assumption, a Mesos/Marathon cluster can be instantiated on the provided cloud resources and the parametrized simulations can run in it as independent containers.

The output carried out from the simulations will be stored in a distributed environment where it can be accessed for post-processing analysis.

6.2 Progress and key results

The MESSAGEix-GLOBIOM integrated assessment model (IAM) relies, via GAMS, on the commercial CPLEX solver, for which applicants have a current academic license. As part of the pilot, applicants have resolved issues related to housing these solvers on EOSC infrastructure. Currently, the license is valid as long as the work is carried out through IIASA. The longer-term solution will be to move towards supporting additional, free/open-source solvers besides CPLEX.

To support this proposal, the following requirements have been addressed:

- **RQ1**: Deploy virtual machines on EGI Cloud Compute: INFN-Bari, vo.iiasa.ac.at: 200 vCPUs cores, 800GB of RAM.

- **RQ2**: Enable federated identity management using one of the available AAI solutions provided by EOSC-hub.
- **RQ3**: Setup the EGI Data Hub for handle 6 TB of distributed storage
- **RQ4:** Setup a Database (PostgreSQL) managed by the applicants
- **RQ5**: Containerization of the application and setup on the VMs a Mesos/Marathon cluster as scheduler for the parametric jobs

A first effort has been made to understand the different components (authentication, registration, VPN, tooling, etc.) needed to access the cloud resources. In addition, an intensive activity is still ongoing to introduce changes to the model code that are necessary to automate job creation, execution and reporting. In this respect, work has been performed by running a few jobs implementing the above mentioned changes and automation and a GITHUB repo has been created to track activities and speed-up communications.

However, while the cloud computing resource was ready to be used by IIASA, the COVID-19 pandemic has heavily influenced the content-related activities foreseen in the work plan for the present EAP application. Key software development personnel moved on to other positions and in the midst of the pandemic it took a very long time to find replacement capacity. Also, key-researcher time needed to be diverted to IPCC obligations, reducing the capability to further develop the MESSAGE model to enable running on the remote computing resource.

New software development capacity has been hired and is supposed to start working with IIASA in June 2021. We have agreed with the service provider to extend all services until Fall of 2021 to enable developing the remote running capabilities of the MESSAGE model and perform an application. The activities will be gradually resumed as soon as the key-personnel will be operational again.

6.3 Service integrations

Service	Provider	Description of the integrations	Allocated ICT resources (cloud, storage, etc.)
EGI Cloud Compute	EGI	Access to the resource was established, but the resource was not used for reason mentioned above	RECAS-BARI: Number of virtual CPU cores: 200 Memory per core (GB): 4 Local disk (GB): 6144 (in total)

			Public IP addresses: yes and a VPN
Online storage	EGI	Access to the resource was established, but the resource was not used for reason mentioned above	RECAS-BARI: Guaranteed storage capacity [TB]: 6 Standard interfaces supported: POSIX, SWIFT1 Storage technology: CEPH, with block storage exposed via POSIX2

6.4 Lessons learnt

The Early Adopter Program helped to familiarize our research community with the existence of the EOSC and how to obtain access to the resources.

6.5 Impact

The services that the pilot provides are of key importance for the modeling capabilities of our community and therefore we are mentioning and promoting it where possible.

6.6 Future plans and sustainability aspects

The service provider has offered to keep the EOSC resources available for IIASA until the fall of 2021. This will allow us to develop the planned modeling capabilities and perform a scientific analysis.

7 Open AiiDAlab platform for cloud computing in Materials Science

Principal investigators: Aliaksandr Yakutovich (EPFL), Giovanni Pizzi (EPFL)

Shepherd: Enol Fernandez (EGI.eu)

7.1 About the pilot - initial ambition

AiiDAlab brings the AiiDA workflow manager for computational science (www.aiida.net) to the cloud. While domain experts can install AiiDA on their own hardware, the AiiDAlab web platform gives novice users access to their personal pre-configured AiiDA environment through a web browser. AiiDA is a workflow manager for computational science with a strong focus on provenance, performance and extensibility. When executing a workflow, AiiDA records the provenance − calculations performed, codes used and data generated − in a directed acyclic graph tailored to provide full reproducibility of any given result.

The goal of the pilot is to deploy AiiDAlab on EOSC-hub resources in the CESNET Czech computing centre, belonging to the EGI Federated Cloud, combined with EGI Check-in for authentication.

7.2 Progress and key results

During the Early Adopter Programme, we were able to put AiiDAlab on the CESNET Kubernetes cluster and connect it with the EGI Check-in instance providing any interested user with free access to AiiDAlab platform. The machine is now capable of unboarding up to 50 concurrent users. To easily connect AiiDAlab to HPC resources we developed a set of tools that are straightforward to use.

The pilot project helped us to increase the technological maturity of the AiiDAlab container making it able to run on Kubernetes cluster. Additionally, we were able to give access to AiiDAlab for a wider audience of users. AiiDAlab is currently listed in the EOSC Portal: https://marketplace.eosc-portal.eu/services/aiida-lab .

7.3 Service integrations

Service	Provider	Description of the integrations	Allocated ICT resources (cloud, storage, etc.)
EGI Cloud Compute	EGI	AiiDAlab-demo instance is deployed on a Kubernetes cluster set on the CESNET computing resources	CESNET: 50 CPUs, 200 GB RAM, 1 TB disk storage.

EGI Check-in	EGI	EGI Check-in is used for registering users and authenticating them for AiiDAlab.	

7.4 Lessons learnt

The most challenging aspects of AiiDAlab@EOSC deployment were ensuring the stability and scalability of the service. From one side, this requires the stability of the Kubernetes cluster, from the other side this introduces some constraints on handling users' resources if one wants to provide access to the platform to a wide audience.

The first aspect, scalability of the Kubernetes cluster, despite expectation is not always granted and requires careful setup. For instance, in our case it took about 8 month for the Kubernetes cluster at CESNET to be fully operational.

The second aspect, handling of user's resources, requires some re-thinking and not yet fully-resolved. For sustainability of the platform, we will have to enable mechanisms for (a) stopping the user's container and (b) deleting the user's data volume when they are not used. In the current implementation a container is stopped when the user didn't connect the service for longer than 48 hours. This approach doesn't cover all the cases, as it can happen that the user has submitted an AiiDA workflow which does not require intervention and might run for longer than 48 hours. Deleting the user's volume might be necessary for the cases when it was used once and remained inactive for a long time. Such cases should be automatically detected and the disc space should be released by the platform.

For the attractiveness of a platform the registration procedure should be as quick and as streamlined as possible. In this respect, the current implementation of the EGI Check-in platform has yet room for improvement. The registration time for an experienced user is currently about 4 minutes and should be lowered as much as possible (desirably, by a factor of 2 or even more). It also takes users through a significant amount of different pages that appear redundant and might be even skipped.

7.5 Impact

As a result of the AiiDAlab pilot project, we have managed to set up an AiiDAlab platform that is open to the whole world. In case the platform will stay up and running, it will be used as a "landing point" for the new users who would like to try AiiDAlab for the first time.

7.6 Future plans and sustainability aspects

AiiDAlab is one of the deliverables of the NCCR MARVEL Open Science Platform (https://nccr-marvel.ch/research/ii/platforms/open-science-platform), which makes computational science more accessible. To ensure the sustainability of AiiDAlab we must maintain good service quality and lower as much as possible the entry barrier for the newcomers. From this perspective, the AiiDAlab machine with free access that is currently deployed at EOSC is very well in line with the sustainability goals.. To achieve a good level of its maturity we will participate in 'Call for Use cases' from the EGI-ACE EOSC project (https://www.egi.eu/projects/egi-ace/call-for-use-cases/).

8 VESPA-Cloud

Principal investigator: Baptiste Cecconi (obspm)

Shepherd: Baptiste Grenier (EGI.eu)

8.1 About the pilot - initial ambition

VESPA (Virtual European Solar and Planetary Access) is a mature project, with 50 VESPA providers distributing open access datasets throughout the world (EU, Japan, USA). In October 2019, the current number of data products available within the VESPA network reaches 18.3 millions (among which 5 millions products from the ESA/PSA, Planetary Science Archive).

The VESPA team is supported by the Europlanet-RI-2024 project (started on Feb 1st 2020 for 48 months, H2020 grant agreement No 871149).

Each VESPA provider (institutes, scientific teams...) is hosting and maintaining a server (physical or virtualized) with the same software distribution (DaCHS, Data Centre Helper Suite), which implements the interoperability layers (from IVOA, International Virtual Observatory Alliance, and VESPA) and following FAIR principles. Each server hosts a table of standardized metadata with URLs to data files or data services. Data files can be hosted by the VESPA provider team, or in an external archive (e.g., ESA/PSA - Planetary Science Archive).

The VESPA architecture relies on the assumption that data provider’s servers are up and running continuously. The VESPA network is distributed but not redundant. For small teams with little or no IT support available locally, the services are down regularly. We thus need a more stable and manageable platform for hosting those services. The EOSC-hub service “EGI Cloud Container Compute” would solve this problem.

We propose to use the EOSC infrastructure to host VESPA provider's servers (through a controlled deployment environment with git-managed containers).

The open-source DaCHS framework is developed for Debian distribution. A docker containerization will be used to facilitate the framework deployment on other Linux environments.

8.2 Progress and key results

The VESPA team has implemented a prototype that proves the feasibility and relevance of the initial proposal. The prototype proposes a workflow based on the deployment of a docker-based container implementing the Astronomy Virtual Observatory framework (DaCHS, Data Centre Helper Suite), together with selected data services. In the course of the pilot project, VESPA has transitioned to a more sustainable service configuration management, using an eduTEAMS managed VO, for managing the access to the services:

- The docker-based workflow prototype with a small test data service is functional;
- The server configuration files are managed on a gitlab server (hosted by Obs. Paris);
- The data services configuration are managed on a gitlab server;
- The access to the gitlab server is managed through eduTEAMS AAI;

- The cloud compute resources were provided by EGI through its resource centers CC-IN2P3 and CESNET;
- The mapping between an “admin:cloud” group defined in EDUTEAMS AAI has been mapped to the VESPA VO at EGI Checkin, to allow the access to the VM deployment;
- The GÉANT team provided the VESPA support for connecting a gitlab server to eduTEAMS.

The VESPA team is willing to consolidate the prototype before onboarding it on the EOSC Portal.

8.3 Service integrations

Service	Provider	Description of the integrations	Allocated ICT resources (cloud, storage, etc.)
eduTEAMS	GEANT	Used as a community AAI to manage the user community's authentication and authorisation	Service access
EGI Check-in	EGI	Used as e-infrastructure AAI proxy, mapping the attributes from eduTEAMS to be consumed by EGI services	Service access and integration with eduTEAMS
EGI Cloud Compute	EGI	Deploying and running the Virtual Machines supporting the service	CC-IN2P3: 10 VM instances linux based, 2 CPU per VM, 4GB RAM per VM, 20 GB disk per VM, 2TB of Object Storage (Swift/autre) CESNET: 10 VM instances linux based, 2 CPU per VM, 4GB RAM per VM, 20 GB disk per VM, 2TB of Object Storage (Swift/autre)

8.4 Lessons learnt

The VESPA project architecture has been consolidated thanks to the EOSC-Hub EAP:

- Configurations of servers and services managed by git, which improves the robustness and sustainability of the framework;
- We tested a federated AAI service, and we think it is relevant for this type of service;
- Implementation of AAI-managed gitlab server;
- Development of openstack VM deployment script with our application, from git-managed configurations (for deployment on EOSC or locally);
- Better understanding of the EOSC ecosystem.

8.5 Impact

VESPA-Cloud is still at a prototype stage, but the project consolidated the overall VESPA framework, and opened up new solutions and opportunities for future VESPA service implementations.

Specifically, for data providers who are not able or not willing to host a VESPA server for a long period of time, we now have a working solution for service deployment, either on EOSC or on local data centres.

8.6 Future plans and sustainability aspects

The VESPA community is willing to continue the VESPA-Cloud pilot project, and explore further the use of EOSC resources for sharing solar system data. The prolongation of the SLA with EGI for the provisioning of the EGI Cloud Compute and Check-in services has been extended until March 2022 under the same conditions of the current EAP. In the future, a cooperation model will have to be decided on, in order to continue the access to the resources.

9 OpenBioMaps data management service for biological sciences and biodiversity conservation

Principal investigator: Miklós Bán (UNIDEB)

Shepherd: Miguel Caballer (UPVLC)

9.1 About the pilot - initial ambition

The OpenBioMaps is used for data management by nature conservation institutes, biodiversity research and citizen science projects. OpenBioMaps provides several services that make day-to-day work with data easier, but it does not provide tools for analyzing the data. In this project, we aimed to develop a background service based on EOSC tools and resources that support the interpretation of data from databases on conservation biology and biodiversity. It is a new service solution that facilitates and generalizes the most common high-computational analysis of data stored in such databases. This feature will be available as an OpenBioMaps module through a new user interface on each OpenBioMaps server, allowing for a seamless and transparent connection between databases and analyzes.

9.2 Progress and key results

- We have developed a new and API interface which is running on Computational servers. This API can process calls from OpenBioMaps projects. Creates a computational package which is running in a docker environment and can send back results.
- We have created a new OpenBioMaps module, which is a user interface for managing data and analysis scripts and can manage connection with Computational servers.
- We have created a Computational server network which comprises four servers, where two are in EOSC.
- We have created a TOSCA document and needed recipes to deploy the OpenBioMaps server using the IM Dashboard in the EOSC environment.
- We have successfully deployed an OpenBioMaps VM instance in the IFCA site using the IM Dashboard.
- This setup is used by three use cases. Two projects using these services and resources for performing Random Forest based analyses for discovering spatial distributions of species based on a large number of environmental factors. In the third project, researchers perform machine learning based analyses to automatically identify bird species on records.

9.3 Service integrations

Service	Provider	Description of the integrations	Allocated ICT resources (cloud, storage, etc.)

Infrastructure Manager	UPV	Create a TOSCA document and needed recipes to deploy the OpenBioMaps server using the IM Dashboard.	
EGI Cloud Compute	EGI	OpenBioMapser VM instance has been deployed in the IFCA site using the IM Dashboard.	IFCA-LCG2: 32 CPUs, 32 GB RAM, 0.2 TB HD
EGI Cloud Compute	EGI	OpenBioMaps Computational server instances has been deployed on Virtual Machines in the IFCA site using the IM Dashboard.	IFCA-LCG2: 2x32 CPUs, 2x48 GB RAM, 2 TB HD

9.4 Lessons learnt

Discovering the enormous potential of the EOSC requires significant time investment. If I hadn't gotten into the EAP program, I probably wouldn't have spent that much time understanding the whole system and understanding some services.

During the development period, a major service outage and system damage was found in the IFCA service, which resulted in the loss of our EOSC VMs. This case highlighted that the service provider's already introduced provider-independent distributed resource network is also an important condition for stable implementation in the case of computing servers, which also significantly reduces financial dependencies. Because of this, instead of simply restoring services, we redesigned the implementation of the computing servers and created a distributed computing network in which our EAP VMs also joined.

The users of the OpenBioMaps community are mostly conservation or biodiversity research professionals who spend very little time discovering and understanding new IT tools. OpenBioMaps, which has been operating for 10 years this year, provides an excellent opportunity to integrate EOSC services and thus we could open up the opportunity to show new tools to the user community through a familiar environment.

9.5 Impact

Maintaining long-term-supported servers in the EOSC network would allow creating a new high-capacity OpenBioMaps node, which could provide an opportunity to connect many new citizen science and science projects, especially new ones that have no connections to those currently running. The OpenBioMaps community currently includes projects in Hungary and Romania, and our server network is only in these two countries, although the user projects have many international connections and the data collected is worldwide. It would be particularly good for the development of OpenBioMaps if new independent projects joined the community, which could probably be facilitated by having a Western European server in the OpenBioMaps network.

Without the servers created during the pilot, the computing server network has a minimal capacity for the time being and cannot accommodate larger computations. Although the lack of computational capacity can be a significant limiting factor in the evaluation of conservation research data. Therefore, the OpenBioMaps' transparent and distributed computational capacity could be a precious potential for nature conservation projects.

9.6 Future plans and sustainability aspects

OpenBioMaps has been in operation for ten years and its capacity is constantly growing, although this capacity is not significant and it is especially difficult to provide publicly available resources for many minor projects, as larger projects usually connect to the network with their own machine capacity. The computing server network established during the pilot is still in an experimental stage, the widespread implementation has not been introduced because of the lack of sufficient stable machine capacity, although the computing servers created in EOSC have been used by several projects during the development period and are still used today.

Because OpenBioMaps does not have a permanent financial background, we cannot use paid services. If it is possible to continue to use the resources used so far in a supported way, we would primarily keep the computing servers within the infrastructures provided by the EOSC.

However, a publicly available OpenBioMaps server would only be launched if we could enter a long-term contract, as the contract, which is to be renewed year after year, does not provide a sufficiently stable background to maintain a public service server. Because, otherwise, we would have to have the right capacity to possibly migrate the entire server.

10 Towards a Global Federated Framework For Open Science Cloud: Three Use Cases

Principal investigator: Hussein Sherief (AASCTC) , Jianhui Li (CNIC-CAS)

Shepherd: Giuseppe La Rocca (EGI.eu)

10.1 About the pilot - initial ambition

The project aimed to allow researchers from Africa and China to use EOSC services to analyse and publish datasets on a federated cloud infrastructure composed by EGI and CNIC CAS resources. Initially, the intention was to set up an initial trial phase of Global Open Science Cloud linking EOSC and CSTCloud for the following three specific use cases:

- Disaster risk: CASEarth provides high resolution (8 m) satellite data and radiation satellite images for the simulation of tsunami, hurricane, earthquakes, typhoons, floods and extreme weather.
- Smart City: ESA and CSTCloud provide high resolution data and sensor data for the city of Shenzhen in Guangzhou province, China.
- Precision Medicine: Beijing Institute of Genomics (BIG) provides datasets for analysing genetic make up of diseases.

The results of the three use cases would give a complete picture of how to proceed with the Global Open Science Cloud and what added values would be obtained by it and what major issues need to be resolved. To support this project, the team was composed of EGI, OpenCoasts and PSNC technical experts, the Academia Sinica of Grid Computing (AS), the CNIC Chinese Academy of Sciences and the Almaahad Almutagadem Specialized Computer Training Center (AASCTC).

10.2 Progress and key results

Only the Taiwan Typhoon use case was supported during the EOSC-hub Early Adopter Programme. The main issue with the Smart City use case was due that the needed software developed at Wuhan University was only installed on isolated servers. Migrating the software to CSTCloud was complicated and very time consuming. For this reason the Smart City use case could not be launched.

With the Precision Medicine use case there were several ethical issues. More specifically, there were no ethical agreements on how to use human genomes. The datasets from animal and plant genomes also had similar agreement related issues. For these reasons, the Precision Medicine use case could not proceed further.

The final use case of Taiwan typhoon simulation was carried out successfully. The Academic Sinica of Grid Computing contributed to perform the simulation of the Taiwan typhoon Sanders of 2015 with the WRF-4DVAR software. The data sets used for the simulation also included the Doppler radar data sets. The satellite data set for radiation from China Satellite Data Center were of 15km resolution. Comparison of the simulation results and observation data sets were in very good accuracy. This proved without any doubt that additional satellite radiation data sets would give added value to GOSC by further improving the accuracy of the simulation. Further details about the outcome of the Disaster Risk use case can be found in the following documents:

- https://dicosbox.twgrid.org/cernbox/index.php/s/EBPCN13zvMmdRRU
- https://dicosbox.twgrid.org/cernbox/index.php/s/C9ygBcuGbvEVP5T

Similarly, the OpenCoasts simulation of the storm surge for Sanders 2015 was in good agreement with the observation data sets. The accuracy of the simulation can be further improved by more data sets from Taiwan gauge measurements.

References:

- https://documents.egi.eu/public/RetrieveFile?docid=3702&filename=rel042-2021_DHA-NEC.pdf&version=1
- https://documents.egi.eu/public/RetrieveFile?docid=3702&filename=screen_capture_satellite_copernicus.jpg&version=1
- https://documents.egi.eu/public/RetrieveFile?docid=3702&filename=screen_capture_taiwan_opencoasts.jpg&version=1

10.3 Service integrations

Service	Provider	Description of the integrations	Allocated ICT resources (cloud, storage, etc.)
OPENCoastS service	LNEC	- Completed the report on the Taiwan modeling. - Finished the integration of the WRF forcing in opencasts. - Developed scripts for remote sensing inundation line and publishing of the layer. - Maintained the storm surge inundation forecast for Taiwan based on GFS atmospheric predictions.	
WRF from the DMCC CC	Academia Sinica (AS)	Performed the simulation of the Taiwan typhoon Sanders of 2015 with the WRF-4DVAR software	

Cloud resources	PSNC	Provided access to the cloud resources	6 VMs with 8 vCPU cores, 128GB of RAM and 100GB HDD of local disk

10.4 Lessons learnt

The main lesson learnt is that more data sets should be used to increase the accuracy in the simulations.

10.5 Impact

The pilot demonstrated that the Global Open Science Cloud (GOSC) can increase the added value of the data processing by providing more accuracy.

10.6 Future plans and sustainability aspects

For the long-term sustainability of this activity after the end of the project, AASCTC has already requested funding from the European Investment Bank and Asian Infrastructure Investment bank for launching of GOSC in Sudan that would be linking EOSC, CSTCloud with Africa Arab Science and Technology Cloud (AAScTCloud)[4].

[4] https://u.pcloud.link/publink/show?code=XZ21vfXZCimMiVP31jRjFA4mKFIM9hst8xuV

11 Supporting FAIR data discoverability in clinical research: providing a global metadata repository (MDR) of clinical study objects

Principal investigator: Sergei Gorianin (ECRIN)

Shepherd: Hans Piggelen, van (SURFsara BV), Stefano Nicotri (INFN)

11.1 About the pilot - initial ambition

Background - The need to make clinical research data and documents FAIR

In recent years there has been a growing acceptance that to accurately assess the results of trials and other clinical research, and in particular to combine the results from different trials in meta-analyses, it is much better to have access to the original source data, the Individual Participant Data (IPD), as well as the result summaries found in published papers.

In addition, to make sure that the IPD can be fully understood and properly analysed, a variety of other study documents (protocols, analysis plans, etc.) are required. As a result, under pressure from funders and journal editors, more and more researchers are making such material (generically, "clinical trial data objects") available for sharing with others. The datasets are rarely freely available - instead a variety of access mechanisms (e.g. individual request and review, membership of pre-authorised groups, or web based self-attestation), are used in combination with different access types (e.g. download versus in-situ perusal). Furthermore, the various data objects are stored in a wide variety of different locations: a rapidly growing number of general and specialised data repositories, trial registries, publications, the original researchers' institutions, etc.

The researcher or reviewer wishing to locate relevant data objects for a study is therefore faced with a bewildering mosaic of possible source locations and access mechanisms, and this problem of "discoverability" will almost certainly become much worse in the future as more and more materials are made available for sharing. Systems are therefore required to make the data and associated documents generated by clinical research more **FAIR**: **F**indable, **A**ccessible, **I**nteroperable, and **R**eusable. The ECRIN Clinical Research Metadata Repository, or MDR, is designed to be one such system.

General - The role of the MDR

The principal role of the MDR is to make the data objects generated from clinical research easier to locate, and to describe how each of those data objects can be accessed, providing direct links to them where that is possible. The central idea is to develop systems that can collect the metadata about the data objects, including object provenance, location and access details, from a variety of source systems (e.g. trial registries, data repositories, bibliographic systems) and aggregate it into a single **MetaData Repository**, the MDR. The system is designed to first assemble the metadata, on a global scale, and using a variety of methods, e.g. files obtained through API calls, direct file downloads, and web scraping (for further details see Data Collection Overview[5]). It then standardises that metadata into a single schema, devised by ECRIN during this EAP to capture the essential information about each object's discoverability, access and provenance (see the ECRIN Metadata Schemas[6]). The MDR then provides access to the standardised metadata through a single system, accessed via a web portal. The portal system carries out comprehensive indexing of the metadata, to support easy searching and filtering, so that researchers can quickly identify the data objects of interest to them.

The system was initially designed and developed within the EU H2020-funded project eXtreme DataCloud (XDC; grant agreement 777367) in a collaboration between ECRIN[7] (the European Clinical Research Infrastructure Network), ONEDATA[8] and INFN (Istituto Nazionale di Fisica Nucleare - Sezione di Bari). The MDR portal was publicly launched on 29 April 2020.

Currently, the MDR instance in production contains about 901.076 data objects and uses as data sources the WHO recognised trial registries, including ClinicalTrials.gov, PubMed, BioLINCC and Yoda. More information regarding the MDR Data Sources[9] can be found on the MDR wiki page.

The first objective within the EOSC-hub pilot was to extend the MDR demonstrator to run in production in the EOSC environment and be part of the EOSC catalogue and to complete the database by integrating all major data sources dedicated to clinical research. The second objective was to include other EOSC services not already included into MDR such as EGI Federated Cloud resources to host the distributed repositories.

Workplan in EOSC-hub

The initial workplan for our pilot within EOSC-hub is provided in the table below and progress towards the specific objectives is summarized in the “Progress and key results” section.

Quarter	Main activities
Q1	• Investigate a new mechanism of ECRIN metadata ‘injection’ and upgrading on OneData environment. • Revision of web-portal. • Investigate metadata schema and requirements for future harvesting by B2FIND
Q2	• Continue the revision of current web-portal, developed within XDC project in collaboration with OneData (web-portal GUI + OneData Environment) and INFN (ElasticSearch + hardware support).

[5] http://ecrin-mdr.online/index.php/Data_Collection_Overview
[6] http://ecrin-mdr.online/index.php/The_ECRIN_Metadata_Schemas
[7] https://www.ecrin.org/
[8] https://www.onedata.org/#/home
[9] http://ecrin-mdr.online/index.php/MDR_Data_Sources

	• Testing and upgrading the web-portal with respect to updated ECRIN requirements. • Continued investigation on harvesting by B2FIND
Q3	• Testing the work produced in Q1 and Q2 • Start to develop ElasticSearch-based APIs in collaboration with INFN • Enable harvesting of a single MDR instance by B2FIND test instance
Q4	• Finalizing the development and integration testing by users. • Support for potential users, including collecting metrics as well as feedback, and feeding back requests for change. • Enable harvesting of one or more MDR instance by B2FIND production instance.

11.2 Progress and key results

Currently the MDR has reached the TRL 8 (system complete and qualified). The pilot achieved the following goals:

- Running in production on EOSC services:
  - ECRIN portal on EGI Data Hub[10]/Onedata service
  - ElasticSearch backend for the platform
- Scalability of the services
- Improved and possibly automated metadata ingestion into ECRIN/Onedata environment
- Updated and integrated ECRIN portal in OneData environment
- Revised ECRIN metadata schema, published in Zenodo (https://zenodo.org/record/4133889#.X_RoY9hKjcs)
- Support for potential users

All the working plan tasks have been completed, which included mainly:

- Revision of the ECRIN metadata schemas;
- Re-injection and re-indexing of the new datasets on the ONEDATA platform;
- Revision of the web-portal user interface with respect to ECRIN requirements, mostly related to provide for users new information which got available in the ECRIN metadata schema v.5;
- Providing ElasticSearch-based APIs;
- Support for potential users, including collecting metrics by integrating Google Analytics in the MDR web-portal. Annex 3 contains an overview of the number of monthly MDR visitors and their per country distribution. Users' feedback on the MDR has been collected and analyzed

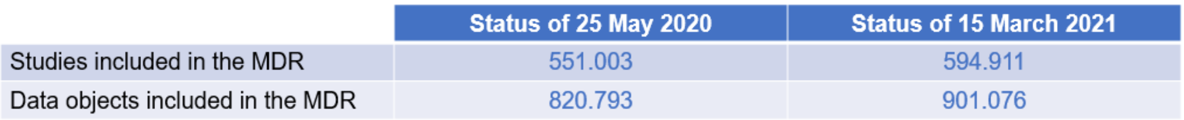
[10] https://www.egi.eu/services/datahub/

through a pilot study. Dedicated contact information is provided in the web-portal to facilitate exchanges with potential users.

The services and resources requested allowed us to extend the scalability of the service running in production with larger hardware resources using EOSC service providers. As a result, we covered the main data sources dedicated to clinical research and provided a user-friendly and efficient platform.

Technical issues encountered with regards to the data injection mechanism and web-portal performance have now been resolved. Searching query execution has noticeably increased within this pilot as well.

The number of studies and data objects available on the web-portal has been increased as well as speed of their importing to the ONEDATA environment. The table below summarizes the status of the MDR in May 2020 vs March 2021 with regards to the number of studies and data objects included.



	Status of 25 May 2020	Status of 15 March 2021
Studies included in the MDR	551.003	594.911
Data objects included in the MDR	820.793	901.076

Our intention is to onboard our services in the EOSC Portal but not at the current stage. Sustainability aspects will need to be clarified beforehand (see section “Future plans and sustainability aspects”).

11.3 Service integrations

Service	Provider	Description of the integrations	Allocated ICT resources (cloud, storage, etc.)
EGI Cloud Compute	EGI	All the described services are deployed on virtual machines hosted on the Cloud@ReCaS-Bari infrastructure.	INFN/RECAS Bari Development web-server: CPU: 8x 2,4 Ghz CPU cores. RAM: 16GB. Disk: 512GB. Production web-server: CPU: 8x 2,4 Ghz CPU cores.

			RAM: 32 GB Disk: 1 TB Database server: CPU: 8x 2,4 Ghz CPU cores. RAM: 32 GB or more Disk: 5TB Testbed server: CPU: 16 VCPUs. RAM: 32 GB or more Disk: 20GB of disk space + 1TB of external storage
ElasticSearch	INFN	ElasticSearch is used as the search engine for the MDR web-portal. It's installed on the servers, provided by INFN, and connected with the ONEDATA system to retrieve and index data from it.	Installed on the servers, provided by EGI/INFN.
EGI DataHub	EGI	The EGI DataHub is a ONEDATA system and has been used for the centralized data storage and as the core platform for the MDR web-portal itself.	Installed on the servers, provided by EGI/INFN.

11.4 Lessons learnt

One lesson learnt during the project was that any direct integration between MDR data and the B2FIND EUDAT service would be difficult to implement and without added value for the entities involved or their users. At an organisational level, it is important to maintain ongoing liaison between ECRIN and its MDR partners and EUDAT and the various services it offers. Possible ways in which the organisations could benefit each other include providing access to relevant B2FIND linked data from the MDR, or registering the MDR as a whole into B2FIND as a resource. Future discussions can also include metadata schemas and APIs.

The pilot within EOSC-hub provided us with access to a variety of trainings and outreach information that were disseminated within the clinical research community and contributed to a better understanding of the current EOSC landscape and its services in the health research field (e.g. catalogue and marketplace and their listed resources).

11.5 Impact

- During this piloting activity, the MDR has been presented in a series of different conferences/seminars/workshops (e.g. PHIRI scientific stakeholder meeting, EGI conference 2020 etc.);
- During the pilot we reached out to different research communities (e.g. pediatric, rare diseases, infectious diseases, public health, modelling). These communities are aware of the services we provide and engaged in dialogue on how we can expand the MDR to fit their specific needs;
- As a result of the pilot, the service provides now a major contribution to the **findability** (F in FAIR) of studies and related data objects in the field of clinical research, covering a wide range of study types, such as interventional trials, observational studies, epidemiological studies based on registries and cohorts. Indirectly, the service also supports **accessibility** (A in FAIR) to data objects in clinical research by providing evidence which objects can be fully assessed and how. The necessity and usefulness of the tool has been highlighted during the pilot through our interaction with the service users. Visibility to and discussion with different research communities (pediatric research, infectious diseases, rare diseases, public health etc.) aimed to increase the synergies between them and avoid duplication of efforts.
- Collection of metrics for the use of the MDR web-portal was implemented in August 2020 with the integration of Google Analytics. Figure 1 shows the number of MDR visitors per month for the period of August 2020 to 15 March 2021. Notably, 308 visitors were recorded for the month of February 2021.

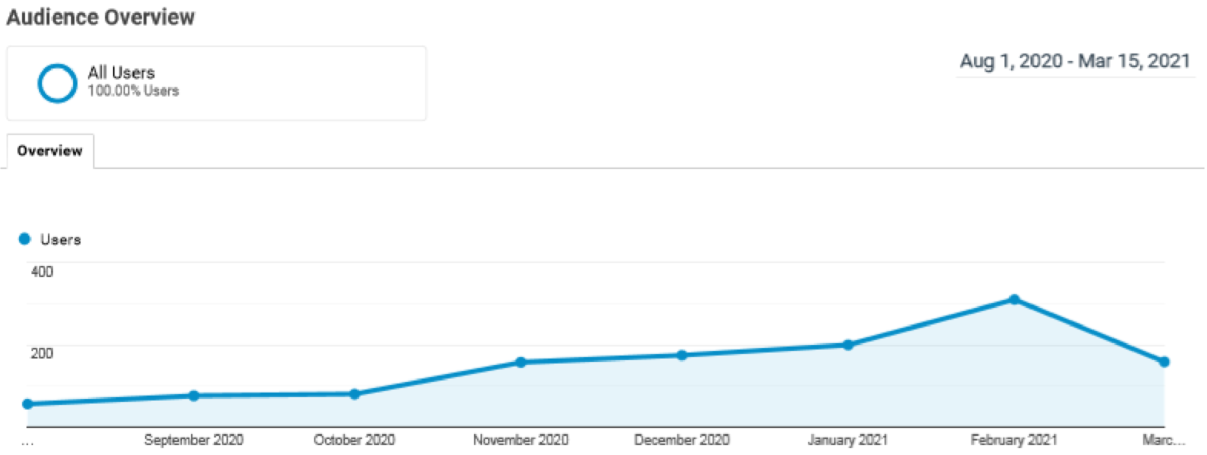


Figure 2. Overview of the number of MDR visitors for the period 1 August 2020 to 15 March 2021 as recorded by Google Analytics.

11.6 Future plans and sustainability aspects

We plan to continue providing the MDR as a service within ECRIN. ECRIN is a sustainable public, non-profit organisation that links scientific partners and networks across Europe (http://www.ecrin.org/) and aims to facilitate multinational clinical research. The current solution covers ECRIN's and communities' needs and requirements, but some extra flexibility of the overall MDR approach may be needed soon due to further collaborations with different projects and other Research Infrastructures and initiatives.

Currently the MDR is using the EGI Datahub/Onedata for centralized data storage and as the core platform for the web-portal and ElasticSearch as the search engine. The system is installed on the servers provided by EGI/INFN.

The next stage for the development of the MDR will take place by participation in different EU projects. For example, within the H2020 EOSC-Life[11] project (grant agreement no 824087) we are expanding the data collection process to include a greater number and variety of data sources, and automate more of the MDR's functioning, making it easier to keep the data as up to date as possible. After the end of EOSC-hub, the data will be collected onto ECRIN's trusted hardware provider and an ECRIN managed portal will be designed. Our intention is to onboard this portal as a service in the EOSC Portal.

[11] https://www.eosc-life.eu/

12 EOSC DevOps framework and virtual infrastructure for ENVRI-FAIR common FAIR data services

Principal investigator: Zhiming Zhao (ENVRI-FAIR)

Shepherd: Andrea Manzi (EGI.eu)

12.1 About the pilot - initial ambition

The project goal is to deploy a DevOps environment, with necessary capacity of Cloud Infrastructures and services, for testing ENVRI-FAIR developments. The project aims to automate the testing/integration of the FAIR data services developed by the teams in ENVRI-FAIR.

The project aims to deliver:

- Automated Cloud execution for data workflow: demonstrate it in the VREs or in ENVRI RIs (e.g., LifeWatch or others). It will help the ENVRI community to learn the EOSC services, and build practices for the other similar use cases;
- Continuously testing and integration for ENVRI services: get familiar with the DevOps/Agile methodologies for software development, testing and operation;
- Notebook based environment for FAIR data access and processing:
    - provide the Jupyter service to users, with examples to access data sets and models,
    - users can perform customised experiments using the notebook services, access data, store the data, publish and share the results with the others.

12.2 Progress and key results

During the project, we developed the following two components:

- FAIR-CELLs, a Jupyter extension to enable the interactive containerization of the Jupyter Cells.
- Cloud-Cells, a Jupyter extension to automate the cloud services (IaaS) provisioning, and container deployment.

We applied these two components in an ecology use case called LidarCloud, in which two legacy programs developed in a previous project are dockerized and executed on remote cloud infrastructures via the Jupyter environment. We run the code on the EOSC IaaS together with the VMs provided by the LifeWatch ERIC. In this way, the original code can be scaled out to process bigger datasets than its original design.

During the project, we have reviewed the DevOps tool provided by Jelastic via the EOSC marketplace. Besides the software engineering support, we investigated the cloud automation support offered by Jelastic. The output of this study has been included in the ENVRI summer and winter schools as part of the training material.

12.3 Service integrations

Service	Provider	Description of the integrations	Allocated ICT resources (cloud, storage, etc.)
EGI Cloud Compute	EGI	VM for Automated Cloud execution for data workflow Jelastic Installation	CESGA: 1 VM 12 CPUs, 16 GB RAM, 1.5 TB storage via NFS 1 VM 8 CPUs , 24 GB RAM , 1 TB storage via NFS 1 VM 12 CPUs, 24 GB RAM, 1.5 TB storage via NFS INFN-CATANIA-STACK: 4 VMs 4 CPUs, 8GB RAM, 100 GB local storage
EGI Notebooks	EGI	EGI Notebooks community deployment	INFN-CATANIA-STACK: 4 VMs 8 CPUs, 16 GB RAM , 120GB local storage

12.4 Lessons learnt

During the EAP, we learned the following lessons:

- We compared our extension with the JupyterHub service offered by EGI, and we did this comparison during the last phase of the project. As a lesson, we think we could have first asked our use case providers to run their code on the JupyterHub environment in the beginning of the project. This could help us better understand the user requirements on the extension we developed.
- We could have better integrated the EAP with the development of ENVRI-FAIR RIs. Due to the late start of the assessment on DevOps tool Jelastic, we did not have enough time to engage development teams from individual RIs.

12.5 Impact

The key outputs of the EAP have been included in the training material of the ENVRI summer and winter school. The technical results have also been presented in the ENVRI-FAIR week. Through those events, the practices learned from the EAP on the EOSC resources have been propagated to the ENVRI community (via WP7 and sub-domain WPs of the ENVRI-FAIR project).

Through the EAP project, the current Laserchicken/Laserfarm components have been expanded from current SURF infrastructure to the EOSC infrastructure in the European scale. In this way, the user communities, and the potential supported data sets will be much broader than the existing service.

12.6 Future plans and sustainability aspects

After the end of the EAP pilot, the technical development and results will be further continued:

- In the ENVRI-FAIR project as part of the knowledge base for supporting developers from ENVRI sub-domains and RIs when developing their data management services;
- In the LifeWatch ERIC as part of the Virtual Lab solutions;
- The developers will exploit the results as part of the EGI Jupyter service, which can be visible in future marketplace;
- We will also actively seek the possible opportunities in the future projects, like EOSC-Future to further sustain the developed solution.

13 Integration of toxicology and risk assessment services into the EOSC marketplace

Principal investigator: Thomas Exner (Edelweiss Connect)

Shepherd: Riccardo Bruno (INFN) , Stefano Nicotri (INFN)

13.1 About the pilot - initial ambition

Chemical risk assessment and the more special nanosafety area have generated large data collections and specialised software tools for analysis, modelling and prediction. OpenRiskNet has started to develop an e-infrastructure and made it available to the communities involved in safety assessment, including toxicology and especially predictive toxicology, systems and structural biology, bioinformatics and its subtopics toxicogenomics, cheminformatics, biophysics and computer science specifically targeting the EU's chemical and nanomaterial manufacturing industries and the corresponding regulatory agencies. At the end of the project in November 2019, 45 services were integrated, that can be grouped into seven categories:

- Toxicology, Chemical Properties and Bioassay Databases,
- Omics Databases,
- Knowledge Bases and Data Mining,
- Ontology Services,
- Processing and Analysis,
- Predictive Toxicology
- Workflows, Visualisation and Reporting

all running on the same core infrastructure.

These services were provided by the OpenRiskNet consortium as well as third parties and include some standard applications like Jupyter tailored to the needs of the communities and integrated with the other services as well as tools for user management and system monitoring. To increase the acceptance of the infrastructure, attract more users and convince follow-up projects to take over the maintenance and further development, OpenRiskNet participated into the EOSC Early Adopter Programme as one of its sustainability measures with three specific goals:

- Make OpenRiskNet more visible by integrating and enlarging the catalogue of available OpenRiskNet services within the EOSC marketplace;
- Allow easy access to the services based on the EOSC AAI;
- Prepare the OpenRiskNet environment to be deployed onto EOSC Cloud Computing and Storage infrastructure.

Since the OpenRiskNet infrastructure had a high technology maturity already at the beginning of the EAP, these tasks are mainly targeting the stronger integration of OpenRiskNet in the EOSC environment and increasing the harmonisation and interoperability with other EOSC services. However, porting the reference instance in form of the publicly available OpenRiskNet virtual environment onto EOSC resources and secure provision of these resources for long-term sustainability was also targeted by the pilot.

In order to achieve the above goals, cloud resources (90 vCPU cores, 200 GB of RAM and 1 TB of disk space) were allocated by EGI in the Cloud@ReCaS-Bari INFN site.

13.2 Progress and key results

Since the pilot just started at the end of the OpenRiskNet project (November 2019), the consortium agreed on a short-term sustainability solution in form of cloud resources provided by one of its partners (Johannes-Gutenberg-Universität Mainz). Even if the infrastructure is not running on EOSC resources, linking to EOSC was achieved in two different ways.

- OpenRiskNet is running a single-sign-on user management system based on KeyCloak. The EGI and ELIXIR AAI were added as additional authentication mechanisms so that EOSC users are able to access the services without the need to create a new user account for OpenRiskNet. In this way, the second goal from above could be achieved without any technical difficulties.
- With the help of the EOSC team, 4 OpenRiskNet services were onboarded in the EOSC portal. This includes the core infrastructure[12] as well as the three specific services Squonk Computational Notebook[13] (a graphical environment to design and execute scientific workflows), Jaqpot[14] and LAZAR[15] (two modelling platforms for generating QSAR and, in the case of Jaqpot, biokinetics models). The OpenRiskNet-internal catalogue was adapted so that all information needed for listing services in the EOSC portal is directly available. However, this does not address the second goal from above completely since an automatic integration of OpenRiskNet service in the EOSC marketplace was pursued to offer the onboarding in the EOSC portal as a service to the OpenRiskNet service providers simultaneously to the listing in the OpenRiskNet catalogue without having to do manual work.

These first successes were followed up with an analysis on possible ways to scale-up the existing OpenRiskNet production site using EOSC ICT resources to be able to serve a continuously increasing user base and react to varying workloads. Multiple EOSC cloud computing services were identified and required changes in the OpenRiskNet infrastructure as well as support services by EAP were discussed. Even the needed EOSC cloud resources were already offered and allocated. However, the running OpenRiskNet reference instance was then heavily influenced by the cyber attack hitting public institutions in the Netherlands and Germany. This caused that all existing human resources available to OpenRiskNet had to concentrate on reestablishing the system to allow existing users to continue their work. This was worsened by the fact that no budget to cover these human costs was available due to the fact that OpenRiskNet had already ended. After the update of the security system done by the University of Mainz to undo the harm caused by the attack and prevent similar attacks in the future, a complete reinstallation of the system was necessary also including an upgrade of all the cloud components and especially the OpenShift system. Due to incompatibilities

[12] https://marketplace.eosc-portal.eu/services/openrisknet-e-infrastructure
[13] https://marketplace.eosc-portal.eu/services/squonk-computational-notebook
[14] https://marketplace.eosc-portal.eu/services/jaqpot
[15] https://marketplace.eosc-portal.eu/services/lazar

in the new and old versions, many OpenRiskNet services had also to be adapted. Shortly after, the Mainz cloud system experienced hardware problems, resulting in the need to redo the reinstallation again.

These unfortunate events showed the immense maintenance costs associated with the system operated by an OpenRiskNet partner and especially with keeping the OpenShift-based infrastructure up to date. To find solutions to reduce these costs and, thus, be able to cover them in the future, the goals of the EAP pilot were changed to support the development of a new sustainability plan for OpenRiskNet technically. This had the two aspects:

- Identify the needed changes to remove the OpenShift components: even if OpenShift offers useful additions to Kubernetes like CI/CD support, the fast update cycles, the missing support of older version and the missing uptake from the cloud community makes it hard to use it in a production instance;
- Identify an at least mid-term solution based on EOSC resources to provide a stable and sustainable OpenRiskNet reference instance (see below).

This was paralleled by discussions with interested user groups and ongoing and planned infrastructure projects to see what the current user base is and how this will project into the future. Additionally this gave a picture on the commitment of the community to sustain the infrastructure. This was needed since, due to the relatively short runtime of the OpenRiskNet projects, focus had to be put on the technical developments and early adopters. Projects like the MSCA ITN in3 and EU-ToxRisk adopting OpenRiskNet functionality and especially NanoCommons and the Dutch VHP4Safety project even committed to support maintenance and continue with the development of the infrastructure demonstrating the clear need for sustaining OpenRiskNet.

13.3 Service integrations

Service	Provider	Description of the integrations	Allocated ICT resources (cloud, storage, etc.)
EGI Check-in	EGI	ELIXIR and EGI AAI are integrated into the OpenRiskNet single-sign-on user management	

13.4 Lessons learnt

Chemical and nanomaterial risk assessment is requiring many specific data management, analysis and modelling services developed in the scientific and neighbouring communities. These specific requirements justify the implementation of a community-specific infrastructure based on the OpenRiskNet core infrastructure and services. Cloud systems are the ideal way to provide such services. However, even if it is important to constantly identify and evaluate state-of-the-art features of these cloud solutions to improve the systems, it requires a large amount of resources to keep it up to date with the fast changing cloud computing field resulting in high maintenance costs if individual systems are provided. The EAP pilot was an excellent opportunity to understand how EOSC services can be used to provide core functionality to the individual community-specific infrastructure freeing it from the need to maintain the cloud system itself, on top of which the specific risk assessment services can be deployed and connected to other EOSC services. In this way, the risk assessment community can concentrate on the harmonisation of data and making tools more interoperable inside the community and across communities in the scientific sense of being able to combine data and tools in a more automated way to solve specific real-world problems.

13.5 Impact

Due to the technical issues with the existing non-EOSC reference instance and the need to remove non-standard components of the core infrastructure to reduce maintenance costs described above, the EAP was only partly successful in better satisfying requirements of the OpenRiskNet users directly. However, by identifying these issues and defining solutions addressing these, the EAP actually had a larger impact than expected by defining a sustainability strategy for OpenRiskNet with much lower maintenance costs as achievable with solutions operated by partners from the scientific community. This helped to convince at least two projects to base their infrastructure developments on the solutions provided by OpenRiskNet and with that on EOSC services. In this way, not only the current state can be sustained but continuous developments, improvements and increase in functionality are secured not only providing better services to the risk assessment / toxicology communities but also for new communities as e.g. represented by users of the NanoCommons infrastructure and partnering projects like the European Material Modelling Council and the virtual human platform of the Dutch VHP4Safety project.

13.6 Future plans and sustainability aspects

With the sustainability plan described above, it is now clear that the OpenRiskNet services can be offered as part of EOSC after the end of EOSC-hub and will be extended and improved to also address needs of neighboring communities. Besides the financial commitments of NanoCommons and VHP4Safety, partners of OpenRiskNet and NanoCommons have been included as associated partners in the EGI-ACE projects, which will provide cloud resources for the next 2.5 years to run the OpenRiskNet reference instance and additional NanoCommons services. These resources will also be accessible to other projects like VHP4Safety either by using the reference instance but also for deploying but still interlinked project-specific instances for developing and integrating their services. The pilot provided the groundwork to now be able to deploy the OpenRiskNet instance on the EGI-ACE resources so that we do not expect any technical difficulties during the deployment.

14 AGINFRA+: Virtual Research Environments to Support Agriculture and Food Research Communities

Principal investigator: Leonardo Candela (ISTI-CNR)

Shepherd: Pablo Orviz (CSIC)

14.1 About the pilot - initial ambition

The goal of the pilot was to rely on the EOSC-Hub resources offering to reinforce the DataMiner cluster serving the AGINFRA+ community. DataMiner is one of the key services forming the Data Analytics part of the AGINFRA+ Platform[16]. DataMiner[17] enacts its users to execute analytics tasks either by relying on methods provided by the user or by others. It is endowed with importing and sharing facilities for analytics methods implemented in heterogeneous forms including R, Java, Phyton, and KNIME. Most importantly from the point of view of the pilot is the fact that it enacts tasks execution by a distributed and hybrid computing infrastructure where computing resources are transparently provided by many providers.

In the very last period of the AGINFRA+ project a Data Science Challenge was organised to make it possible for Startups and SMEs to exploit the innovative solutions produced by the project. BioCoS[18], a Bioinformatics & Biotech company operating in Greece and solving food fraud using DNA, was selected as the winner of the challenge and had the opportunity to showcase its work over the AGINFRAplus software tools and services. More specifically, they were provided with access to the online computing environment of the project, in order to test if their computationally demanding tasks can be executed in a faster and easier way.

In order to serve this use case, a dedicated Virtual Research Environment was created[19] to provide the BioCos team with the services and facilities developed by AGINFRA+ including the DataMiner service. The computing capacity of this environment was planned the be reinforced by exploiting the resources acquired by the pilot.

14.2 Progress and key results

The primary goal of the pilot was to showcase how simple it is to enlarge the computing capacity made available to communities by relying on diverse providers (by EOSC-hub in this case). As soon as the pilot managed to allocate the computing resources (made available by Institute of Physics of Cantabria (IFCA)) a DataMiner cluster was created and allocated to the Virtual Research

[16] Assante, M, Boizet, A, Candela, L, et al. Realizing virtual research environments for the agri-food community: The AGINFRA PLUS experience. Concurrency Computat Pract Exper. 2020;e6087. https://doi.org/10.1002/cpe.6087

[17] Assante M, Candela L, Castelli D, et al. Enacting open science by D4Science. Future Generat Comput Syst. 2019;101:555-563. https://doi.org/10.1016/j.future.2019.05.063

[18] https://www.biocos.gr/

[19] https://aginfra.d4science.org/web/aginfraplus4biocos/

Environment dedicated to the BioCos company. The integration was technically smooth and completely achieved, the exploitation of the resources was very limited for reasons not depending on the overall solution.

The resources the pilot was provided with enabled to create a cluster of 2 DataMiner worker (each with 16VCPU, 60GB Disk, and 29.3GB RAM), this capacity is sufficient to showcase the integration.

14.3 Service integrations

Service	Provider	Description of the integrations	Allocated ICT resources (cloud, storage, etc.)
EGI Cloud Compute	EGI	DM cluster deployment in the EGI Cloud Compute service	IFCA (Openstack 4 VMs, 70 VCPUs, 136.7GB, 2TB Volume)

14.4 Lessons learnt

D4Science[20] (the provider of the DataMiner service) is exploiting resources from several providers including EGI (there is an SLA agreement in place since 2017). This pilot demonstrated one time more how the solutions developed and operated are solid and make it possible to easily enlarge the set of providers.

However, the number of resources made available by the pilot was (as expected) limited with respect to both “size” and “time”. This limitation, especially in scenarios like those supported by D4Science that is used to support heterogeneous communities with a rich array of production level services, poses questions with respect to the ratio between costs and benefits. In the specific case, effort was spent to deploy the cluster on the new resources acquired by the pilot yet when the BioCos community completed the testing phase (demonstrating that their pipeline can be successfully executed on third party resources) but it was impossible to allocate the cluster to serve other scenarios because it is very “small” with respect to those D4Science is usually offering (an average production cluster consists of 15 DataMiner services each with 16 VCPUs, 32GB RAM, and 100 GB Disk Space plus 15 DataMiner workers each with 16 VCPUs, 32GB RAM, and 100 GB Disk Space plus a dedicated load balancer).

14.5 Impact

D4Science is a well-established service provider supporting several communities (more than 150 active Virtual Research Environments made available by 20 Gateways)[21]. The pilot supported one of these Virtual Research Environments and showcased how simple might be to exploit resources

[20] http://www.d4science.org

[21] https://services.d4science.org/thematic-gateways

provided by third party service providers to reinforce the capacity of the working environments D4Science can create to serve the needs of communities of practice.

This pilot demonstrates how the resources operated by diverse service providers could be conveniently mobilised to provide final users with feature rich and user-friendly working environments hiding the complexity of the underlying settings and the fragmentation of the providers.

14.6 Future plans and sustainability aspects

The pilot stems from a use case having a limited duration and scope (i.e. to allow a company to test the services and solutions stemming from a research project). Because of this no future plan aiming at maintaining the resulting environment (the overall Virtual Research Environment) active after the planned deadline (December 2022) was envisaged.

However, the SLA established between EGI and D4Science will continue to be active implying that D4Science will continue to rely on resources and services provided by EGI.