EOSC-SYNERGY

EU DELIVERABLE D4.2

D4.2 - First prototype of the EOSC Thematic services

(DEMONSTRATION)

Document Identifier:	EOSC-SYNERGY-MS17
Date:	31/12/2020
Due Date:	31/12/2020
Activity:	WP4
Lead Partner:	UPV
Document Status:	Final
Dissemination Level:	PUBLIC
Document Link:	http://dx.doi.org/10.20350/digitalCSIC/12610

Abstract:

This document describes the thematic services that expose the applications and data to the scientific community. This is a report that supports the demonstration deliverable D4.2. The 10 thematic services of EOSC-SYNERGY are reasonably different in maturity, requirements, technological needs and approach, which guarantees that the expansion of the EOSC capacity addresses multiple dimensions and challenges. The document describes four of them more in detail, as planned in the DoA.

EOSC-SYNERGY receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 857647

I. Copyright Notice

II. Delivery Slip

	Name	Partner/Activity	Date
From	Ignacio Blanquer	UPV/WP4	
Reviewed by	**Moderator:** Isabel Campos **Reviewers:** Lara Lloret Marcin Plociennik	CSIC/WP1 CSIC external PSNC/WP6	
Approved by	PMB	PO	

III. Document Log

Issue	Date	Comment	Author/Partner
v0.1	4/12/2020	TOC and initial draft version	Ignacio Blanquer / UPV
v0.2	18/12/2020	Sections 3.x.1-3 completed, section 3.x.4-5 for SAPS	Amanda Calatrava /UPV, Alberto Azevedo / LNEC, Juan Sánchez-Ferrero /INDRA, Manuel Pavesio / INDRA, José María Fernández / BSC, Valentin Kozlov / KIT, Laura del Caño / CNB, Jan Astalos / IISAS, Ales Krenek / CESNET, Tobias Kerzenmacher / KIT,
v0.3	23/12/2020	Sections 3.x.4 and sections 3.x.5 for WORSICA, GCore and O3AS	Amanda Calatrava /UPV, Alberto Azevedo / LNEC, Juan Sánchez-Ferrero /INDRA, Manuel Pavesio / INDRA, José María Fernández / BSC, Valentin Kozlov / KIT, Laura del Caño / CNB, Jan Astalos / IISAS, Ales Krenek / CESNET, Tobias Kerzenmacher / KIT,
v0.4	28/12/2020	Sections 3.x.5 MSWSS, SDS-WAS, SCIPION, OpenEBench	José María Fernández / BSC, Antonio Rubio / CIEMAT, Laura del Caño / CNB, Jan Astalos / IISAS
v0.5	31/12/2020	Several sections	Isabel Campos / CSIC, Marcin Plociennik /PSNC

IV. List of Acronyms

Acronym	Description
AAI	Authentication and Authorisation Infrastructure
AEMET	Spanish State Meteorological Agency
AERONET	AErosol RObotic NETwork
B2FIND	EuDat Discovery service based on Metadata
B2SAFE	EuDat Service for distributing and storing large volumes of data
B2STAGE	EuDat service for data ingestion
CCMI	Chemistry-Climate Model Initiative
CEDA	Natural Environment Research Council's Data Repository for Atmospheric Science and Earth Observation
CF	Climate and Forecast
CORSIKA	COsmic Ray SImulations for KAscade
CS	Consortium Spatial Information
CSW	Catalog Service Web
DEM	Digital Elevation Model
DIRAC4EGI	Distributed Infrastructure with Remote Agent Control for the European Grid Initiative
DMP	Data Management Plans
DOI	Digital Object Identifier
DREAM	Dialogue on Reverse Engineering Assessment and Methods
DYNAFED	Dynamic Federations system
ebRIM	Registry Information Model
EC3	Elastic Compute Clusters in the Cloud
EGI	European Grid Initiative
EIRENE	European Environmental Exposure Assessment Network
ELIXIR	Life Sciences ESFRI
EMODNET	The European Marine Observation and Data Network
EMPIAR	Electron Microscopy Public Image Archive
EOSC	European Open Science Cloud
EPA	Environmental Protection Agency's

EPANET	Water distribution system modeling software package from the United States EPA
ERIC	European Research Infrastructure Consortium
ESFRI	European Strategy Forum on Research and Innovation
EuDat	Collaborative Data Infrastructure for Data Preservation
FAIR	Findable, Accessible, Interoperable, Reusable
G-CORE	Earth observation data processing software from INDRA
GA4GH	Global Alliance for Genomics and Health
GEANT4	Toolkit for the simulation of the passage of particles through matter
GEE	Google Earth Engine
GPU	Graphics Processing Unit
HDF	Hierarchical Data Format
I2PC	Instruct Image Processing Center
IdP	Identity Providers
IGAC	International Global Atmopsheric Chemistry
IM	Infrastructure Manager
INGENIO	Spanish Earth Observation Satellite
INSTRUCT	Integrated Structural Biology Infrastructure
JSON	JavaScript Object Notation
LAGO	Latin American Giant Observatory
LANDSAT	Earth Resources Technology Satellite
LSDF	Large Scale Data Facility
LSDMA	Large-Scale Data Management and Analysis
MODIS	Moderate Resolution Imaging Spectroradiometer
MSWSS	Modelling Service for Water Supply Systems
NAMEE	Northern Africa, Middle East and Europe
NASA	National Aeronautics and Space Administration
NCEI	National Centers for Environment Information
netCDF	Network Common Data Form
netCDF	Network Common Data Form
NWP	Numerical Weather Prediction
O3AS	Ozone (O3) Assessment
OGC	Open Geospatial Consortium
OGC SOS	Open Geospatial Consortium Sensor Observation Service

OneData	Distributed Data Management solution from Cyfronet
OPENCoastS	Coastal circulation on-demand forecast
OpenEBench	Benchmarking service for Bioinformatics from ELIXIR
PAZ	Spanish Earth observation and reconnaissance satellite
PDGS	Payload Data Ground Segment
POSIX	Portable Operating System Interface for X
QFO	Quest for Orthologs
RECETOX	Research Centre for Toxic Compounds in the Environment at Masaryk University
ROOT	Data Analysis Framework from CERN
SAPS	Serviço Automático de Processamento do SEBAL
Scipion	Cryo em image processing framework. Integration, traceability and analysis
SDS-WAS	Sand and Dust Storms Warning Advisory and Assessment System
SEBAL	Surface Energy Balance Algorithm for Land
SGE	Sun Grid Engine
SIC	Satellite Imaging Corporation
SMOS	Soil Moisture Ocean Salinity
SPARC	Stratosphere-troposphere Processes and their Role in Climate
TCGA	Cancer Genome Atlas
UAV	Unmanned Aerial Vehicles
UMSA	Untargeted Mass-spectrometry Analysis
WCD	water-Cherenkov detectors
WebDav	Web Distributed Authoring and Versioning
WMO	World Meteorological Organisation
WORSICA	Water mOnitoRing SentInel Cloud plAtform
ZBGIS	Basic Slovak database for GIS
Zenodo	OpenAIRE repository for Open Science

Table of Contents

Executive Summary

EOSC-SYNERGY aims at expanding the uptake of EOSC by building capacities. Thematic services constitute an important part of EOSC-SYNERGY and are the final layer that is exposed to final users. EOSC-SYNERGY has identified ten thematic services addressing four scientific areas (Earth Observation, Environment, Biomedicine and Astrophysics). Those thematic services gather and expose data and processing services directly to researchers in a convenient interface.

The thematic services are evolving in EOSC-SYNERGY by refactoring its architecture and integrating EOSC services from the EOSC marketplace. This will lead to increased performance and capacity as well as to enhance its functionality.

The ten thematic services are in operation. They have released updated versions with not the full functionality, but including some innovations performed in the frame of EOSC-SYNERGY. Those thematic services have different and complementary requirements and cover the full spectrum of the integration with the key technical services selected (Authentication and Authorization via Check-in and Life Sciences AAI, cloud orchestration through Infrastructure Manager, Elastic batch queues and Kubernetes through EC3 and access to distributed storage through Dataverse and B2SHARE).

The ten thematic services also constitute useful best practices for future new services to be developed, as they address challenges on metadata management, elastic data processing, interoperability with data infrastructures, federated AAI and accounting, that are common for many scientific domains. The new architecture of the thematic services incorporate seamless integration of the Authentication and Authorization for resources, processing and data as well as the embedded resource provisioning and elastic resizing of resources according to workload and efficient access to data storage. Further evolution will improve and consolidate the existing functionality.

The ten thematic services have released a first version and have a clear plan for adopting the further improvements that are already considered in their plan. The thematic services have produced demonstration videos that have been uploaded to the EOSC-SYNERGY YouTube channel https://www.youtube.com/channel/UC32yLKlcngqrkc791cguUFg.

1. Introduction

This report belongs to WP4, “Capacity building for Thematic Services”. This activity aims at expanding the capacity and capabilities of ten thematic services identified in the project. These thematic services have been partially redesigned and adapted to leverage the functionality offered by services in the EOSC marketplace aligning their architecture to the other services in EOSC.

This document describes the status of the ten services, including demonstration videos.

1.1. Scope of the document

This deliverable is of type “DEM” which refers to pilots, prototypes or demonstrators. This document is a short report that organizes and describes the pilot prototypes developed which constitute the actual deliverable. The document should be considered as a guideline to understand the scope of the services and to evaluate the adaptation performed. Demonstrations are included in the form of videos that outline the advantages of the improvements using EOSC services.

1.2. Target Audience

This document serves the project partners as a summary of the actual progress of the thematic services, as well as an updated description of the architecture and EOSC services involved. This information is relevant for both WP2 (for the provision of services) and WP3 (for the identification of key services whose quality should be evaluated). Finally, this document will serve the evaluators of EOSC-SYNERGY to evaluate the progress of the action with respect to the metrics defined.

1.3. Structure of the document

In addition to this introduction, this document is structured in four main sections. First, section 2 describes in general the service adoption plan for the ten thematic services. Section 3 describes the ten prototype cases and serves as a verification mean for milestone MS18 “All prototype services integrated”, which states in the DoA that all the thematic services have been integrated in the EOSC and offer (but limited) functionality in production mode. Finally, section 4 draws up the conclusions.

2. Summary of the Thematic Services

The ten thematic services have complementary requirements and features. However, in general they share needs on four different categories:

- **Authentication and Authorization Infrastructure** (AAI). All cases require users to be authenticated and authorised. In some cases, there is a need for delegation from the users that access the platform for accessing data or processing resources. In those cases, it is mandatory to have a coherent single-sign on mechanism. Other cases may require an AAI linked to popular scientific IdPs and implement the authentication via Virtual Organization membership.
- **Workload Management**. Most of the cases deal with the execution of a set of batch jobs. In those cases, workload managers should be integrated. This will provide the capability to deal with a larger capacity. Options range from using a standard batch queue (SLURM) eventually powered up with automatic elasticity to using Kubernetes for the orchestration of containers.
- **Resource Management.** Most of the thematic services require deploying a virtual infrastructure where the services that provide the functionality and the processing will take place. In most cases, the use of Infrastructure Manager (IM) or Elastic Compute Clusters in the Cloud (EC3) could provide the capability of defining a virtual infrastructure as code and deploying it on the cloud.
- **Data Storage**. The services need to have a storage connected to the processing that can be efficiently accessed. In this case, there is a wide range of different solutions, ranging from EGI-DataHub and B2Share to local solutions based on Nextcloud, Datavers, Elasticsearch and WebDav.

Figure 1 shows the thematic services and the technology solutions.

Service	*WORSICA*	*G-Core*	*SAPS*	*Scipion*	*LAGO*	*SDS-WAS*	*UMSA*	*MSWSS*	*O3AS*	*OpenE Bench*
AAI	***EGI Check in***	Kerberos LDAP & CAS User/pwd	***EGI Check in***	***EGI Check in***	***eduTEAMS + EGI Check-in***	***B2ACCESS***	***EGI Check in*** & Life-science AAI	***EGI Check in***	***EGI Check in***	Life Sciences AAI
Workload Mng.	ArcCE, Batch (SLURM)	GCore+ K8s	K8s	Batch (SLURM)	Batch (SLURM)	Batch (SLURM)	Batch (SLURM) in ***IM/EC3*** (in Galaxy)	Batch (SLURM) in **EC3** (in Galaxy)	Cluster batch (SLURM) & K8s	GA4GH WES/TES stack + NextFlow
Resource Mng.	***IM (TOSCA)***	***IM / EC3***	***IM / EC3***	***IM / EC3***	***Local clusters & IM+EC3***	***EC3***	***IM / EC3***	***IM / EC3***	**IM**	one
Data Storage	Nextcloud, Dataverse	ElasticSearch for the catalogue	OpenStack Swift	Local + S3	***EGI DataHub ONEDATA***	***B2HANDLE /B2SAFE***	Local + S3	Local + ***ONEDATA***	WebDAV	Local + ***B2SHARE***

Figure 1: Services to manage each one of the four functionality blocks for each thematic service. Green denotes that they have been implemented at the moment of the deliverable and grey boxes are functionality that will be included along 2021. Underlined services are services listed in the EOSC marketplace.

3. Thematic Services

3.1. WORSICA - Water Monitoring Sentinel Cloud Platform

3.1.1. Description

WORSICA is a service for the detection of water using satellites, Unmanned Aerial Vehicles and in-situ data. The main products of the service are: i) coastline detection, which includes coastal inundation areas due to storm-surge events; ii) inland water bodies detection, such as lakes, reservoirs or dams; and iii) water leaks detection on irrigation networks. This thematic service aims at integrating multiple-source remote sensing and in-situ data to determine the presence of water in coastal and inland areas. It is applicable to a range of purposes, from the determination of flooded areas (from rainfall, storms, hurricanes or tsunamis) to the detection of large water leaks in major water distribution networks. It builds on components developed in both national and European projects, integrated to provide a one-stop-shop service for remote sensing information, integrating data from both the Copernicus satellite and drone/unmanned aerial vehicles, validated by existing online in-situ data. The WORSICA service will be available without cost to all european public research groups. The private sector will be able to use the service, but some usage costs may be applied, dependending on the type of resources needed by each application/user.

The integration of the WORSICA service in the EOSC infrastructure will boost the usage of the service at an European level. This service will enable the research communities to generate maps of water presence and water delimitation lines in coastal and inland regions. These products can be useful for emergency and planning methodologies in case of inundations or reservoir leaks. In particular, the service promotes 1) the preservation of lives during an emergency, supporting emergency rescue operations of people in dangerously inundated areas, and 2) the efficient management of water resources targeting water saving in drought-prone areas.

3.1.2. Architecture

The architecture of WORSICA consists of three core components (Fig. 2): i) a frontend component; ii) an intermediate component; and iii) a processing component. The frontend component manages all the interaction of the service with the users through a web portal, such as the configuration of the simulations and requests for the service. The intermediate component is a task orchestrator that manages all the requests that arrive from the frontend and sends them to a processing component, and also deals with input/output storage tasks, such as download/upload of the satellite images and intermediate products and metadata to be sent to the Dataverse repository. The processing component (in purple) is a container with the requirements for image processing (scripts, inputs and software) and sent to be run by a resource manager on the cloud/grid infrastructure.

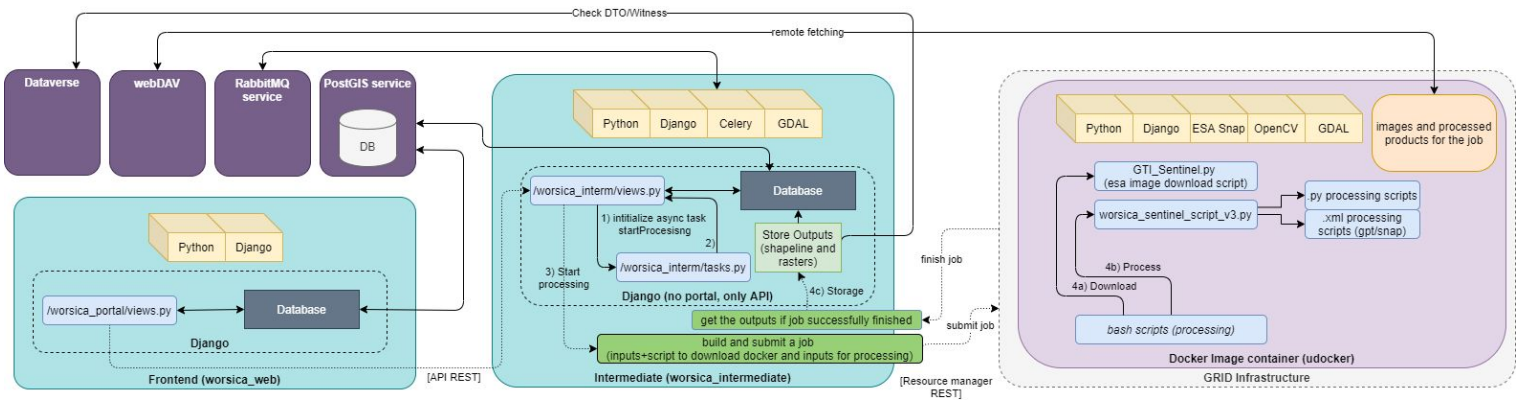


Figure 2 - Architecture of WORSICA service.

3.1.3. EOSC Services

In the EOSC Synergy, technical aspects of the WORSICA service are being improved considerably, using advanced technologies such as high-performance computing and cloud. The service is being scaled up to a European level to reach all interested research communities. We have adapted the service to other European Open Science Cloud (EOSC) services, such as EGI Check-In, and include these in our workflow, such as Dataverse. Therefore, several IT services, available in the EOSC marketplace, are being implemented in WORSICA service.

- **Authentication:** WORSICA uses **EGI Check-In** for the user authentication to the Frontend (portal), and this is a requirement in order to use the available EOSC services.
- **Workload Managers:** Processing jobs submissions are sent by the WORSICA Intermediate service to a GRID infrastructure by using **ArcCE with SLURM.** This allows efficient management of the available resources for HPC in order to speed up the processing jobs.
- **Data Manager: Nextcloud** is used to store processed job submission data input/outputs. **Dataverse** is used to register processed job submission metadata information for data FAIRsFAIR compliance.
- **Ansible and IM tools**: IM is used to deploy the infrastructure required for job processing, repositories and microservices. SLURM and Kubernetes clusters are deployed using TOSCA template over IaaS service and the remaining services will be installed from Docker images. Configurations for SLURM and Kubernetes are set up by ansible playbooks. This will be implemented in the milestone MS18 (All prototype services integrated).

3.1.4. Service Endpoint

The WORSICA web interface manages all the communication with the users. The portal can be accessed using the EGI federated authentication (figure 3) or simply with a verified email.

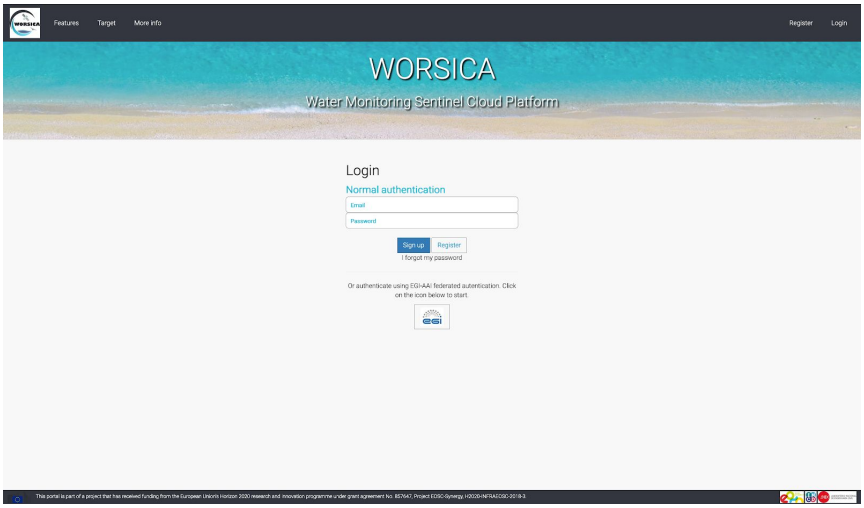
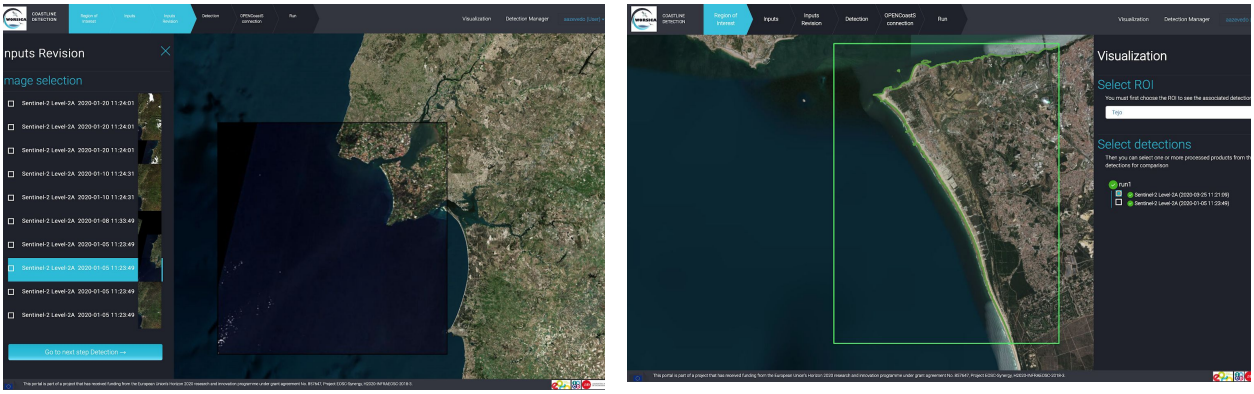


Figure 3 - Login page for the WORSICA service.

In the WORSICA's portal the user can follow a configuration workflow (upper left row, figure 4-a). The main workflow consists in the following procedure: i) user selects the Region of Interest (ROI); ii) the user chooses the type of images to be processed (Sentinel-2, Pleiades or Drone) and the monitoring period; iii) The user verifies which images should be processed; iv) In the Detection menu, the user can select the water index, the number of classes for the clustering procedure; v) afterwards, the user can specify the informations for the connection to the OPENCoastS service, in order to retrieve the tidal elevation for the same period of the images; vi) in the last step, the user can confirm all the configuration values and submit the simulation. After the submission of the simulation, the user can check the results on the visualization menu (figure 4-b)



a) b)

Figure 4 - Snapshots of the WORSICA portal. a) Selection of the images that will be processed; b) presentation of the results of the coastline detection product.

3.1.5. Demonstration Video

The demonstration vídeo prepared for the present deliverable shows briefly the main implementations made to the WORSICA service during the EOSC-SYNERGY project.

The video is divided into two main parts. The first part presents the WORSICA service, the main products of service, the new architecture, the EOSC IT services implemented, and the main achievements obtained with the EOSC-SYNERGY project. The second part is an application of the service for the product of Coastline Detection, where one can see the prototype of the web portal in action and in quasi-real time, with all the IT services from EOSC already implemented. The video can be found at the following link: https://youtu.be/957m5bELN8Q

3.2. G-CORE

3.2.1. Description

G-CORE is a production-ready technology used as a service at ESA's and national programs led by INDRA for the acquisition, storage, cataloguing and processing data from several Earth Observing System (EOS) missions. G-CORE provides two main functionalities:

- A Data Manager for spatial and non-spatial purposes.
- A Processing framework to host external processors developed by third parties to generate added value products based on Satellite imageries.

The objective of the adaptation of the thematic service is to explore the sustainability of the EOS services exposed through the creation of added-value products through the integration of G-CORE as a data manager.

With this in mind, the G-CORE cloud capabilities will provide a processing environment with capabilities for deploying processing prototypes following the SaaS models, without investing in dedicated hardware resources. The ability to deploy new processing frameworks will allow external users to conveniently deploy them to validate their developments.

It means that the G-CORE can be offered as a Payload Data Ground Segment (PDGS) in the cloud for future ground segment space missions to be implemented with the dedicated modifications for each mission, or as a processing framework to plug in different processors that can make use of the Copernicus resources or private data in order to produce different levels of products to be delivered to the users.

3.2.2. Architecture

As previously mentioned, the G-CORE is defined as a system that fulfils most of the Ground Segment needs related to the creation of a scalable and elastic processing systems, including the capability to manage distributed and multiplatform deployments of all their components.

The G-CORE is composed of a set of common components that are able to deploy different instances that will offer the functionality of the components that compose a Ground Segment (specially indicated for processing systems).

Figure 5 represents a typical GCORE structure with a marketplace where different services are offered from third parties. The GCORE will offer different GCORE Instances types that can be deployed according to the user requests. Each user can request a specific service that implies the deployment of a GCORE Instance to perform the activities for the services. The capability of GCORE to integrate different processors and to deliver its results allows creating new services in an easy way.

In addition, all services available would benefit from the elastic and scalable up and down capacity of GCORE in order to reduce the operational costs of a third-party project.

This framework will be offered as a component to be included in complex system to implement the processing chain or as a processing framework to create hybrid infrastructures for processing activities on demand or could be published as SaaS to publish the services in the market place to be used by companies/entities or institutions focused to create added value services/products over satellite imageries.

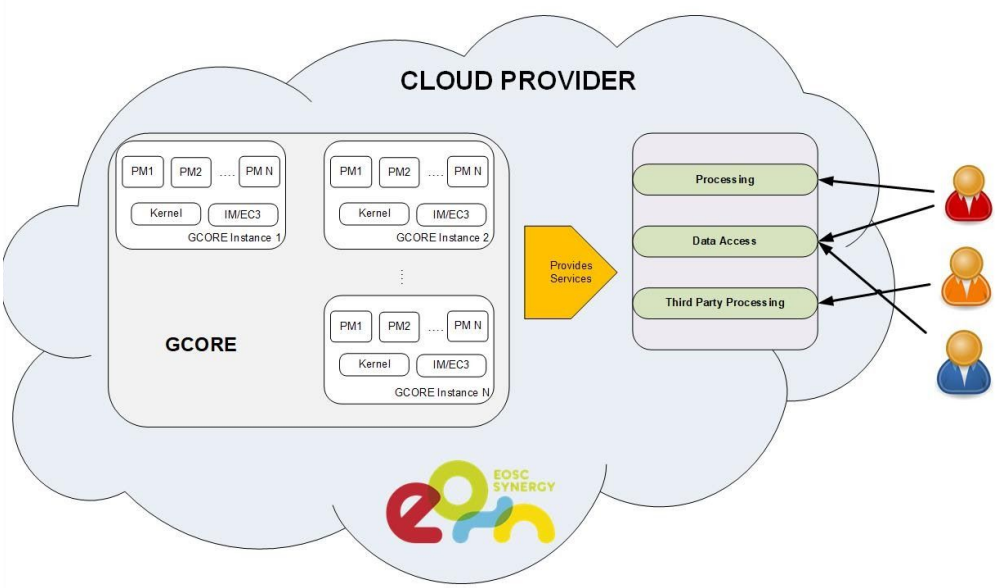


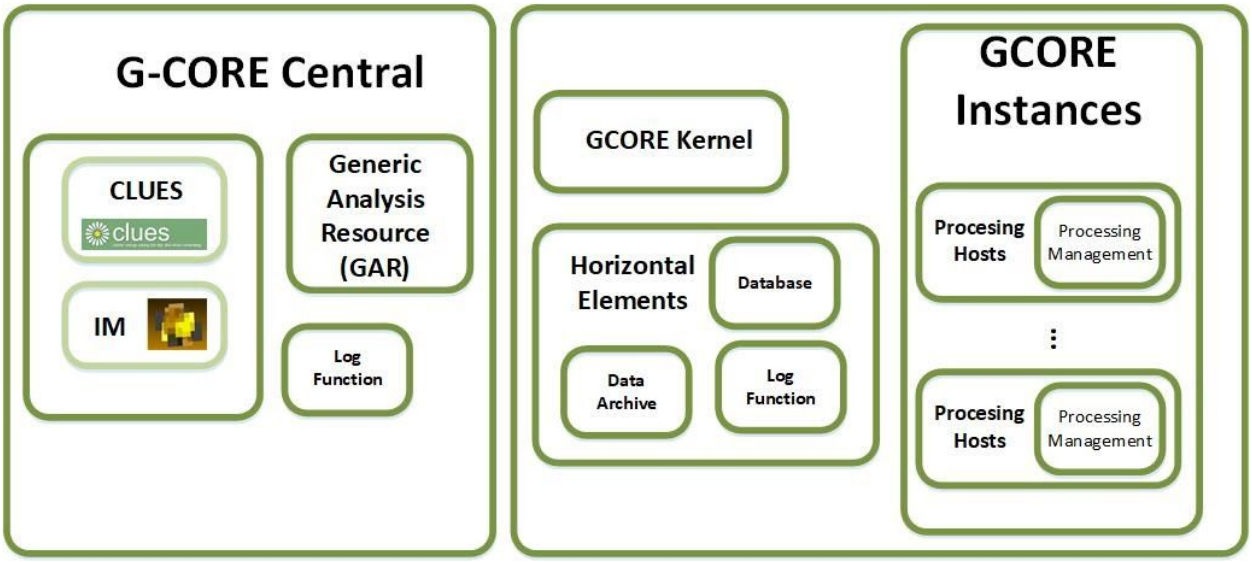
Figure 5. High level architecture for GCORE as EO thematic services

The G-CORE is composed of a set of common components that are able to deploy different instances that will offer the functionality of the components that compose a Ground Segment (specially indicated for processing systems).

In this way, the deployment of the CDPS, shown in figure 6, is based in the following elements:

- G-CORE Central that compose the common elements necessary to support the automatic deployment of the GCORE Instances and that is composed by:
  - An Infrastructure Manager service: composed by the Infrastructure Manager (IM) and CLUES provided by the UPV that allows to deploy and contextualize a set of VM instances and containers in different cloud providers that allows to deploy the G-CORE Instances.
  - Generic Analysis Resources: This component is used to configure the deployment and auto-deployment features depending on the user needs. Also, it includes the monitoring of the jobs deployed.
  - Log System: System in charge to receive the logs of the system in order to show to the operator an unique entry point.
- GCORE Instances that include the deployment of a specific function that a specific project requires. Each instance is composed by:

- Kernel services: The Kernel micro-services include all services associated to the GCORE that compose the core and that enable the basic functions
- Processing Services: These are a set of services that allow the execution of different processes or processors in order to transform a specific input into an elaborated output.



Ffigure 6. High level architecture components for GCORE

3.2.3. EOSC Services

The G-Core service targets the following three user profiles:

- EO data for the science community to use the satellite data in the scientific studies.
- EO data for public organizations to use the satellite imageries as background data.
- EO data for value adders to create added value products from satellite images.
- It will help to define new products and services mixing Earth Observation data with other types of data for scientific and social environments

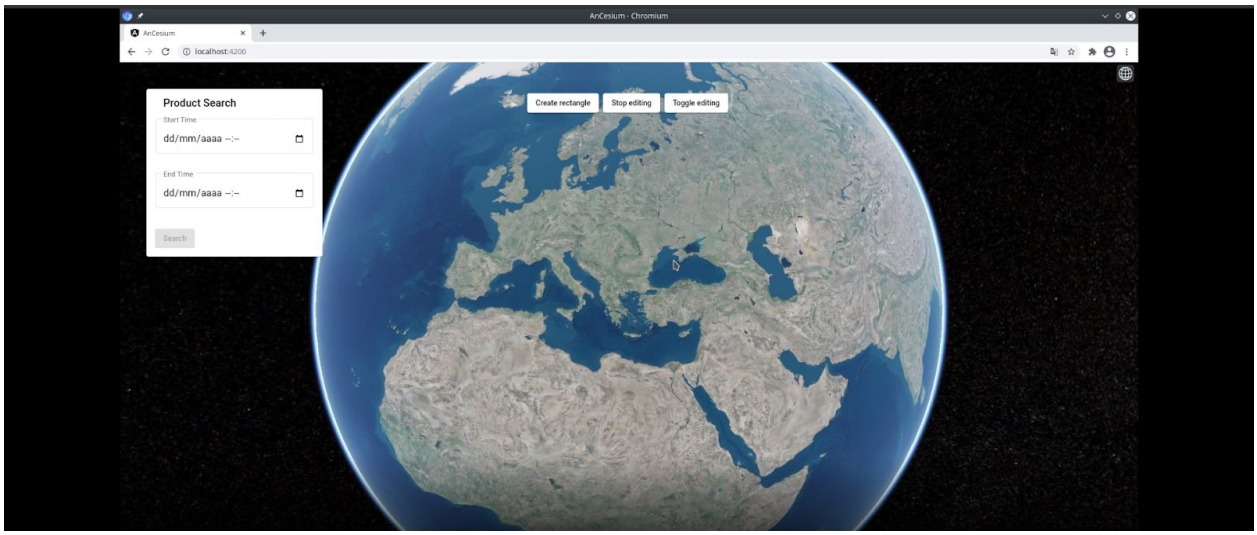
The expected impact of the adaptation of the service is to democratize the usage of EO data out of the scope of nominal fields. It will help to define new products and services mixing Earth Observation data with other types of data for scientific and social environments.

In addition to the proposed services G-CORE makes use of the next common services:

- EGI check-in for authentication. G-CORE makes use of this method in addition to its own login and authentication methodology based in a platform-SSO.
- IM and CLUES. Composed by the Infrastructure Manager (IM) and CLUES provided by the UPV that allows to deploy and contextualize a set of VM instances and containers in different cloud providers that allows to deploy the GCORE Instances.

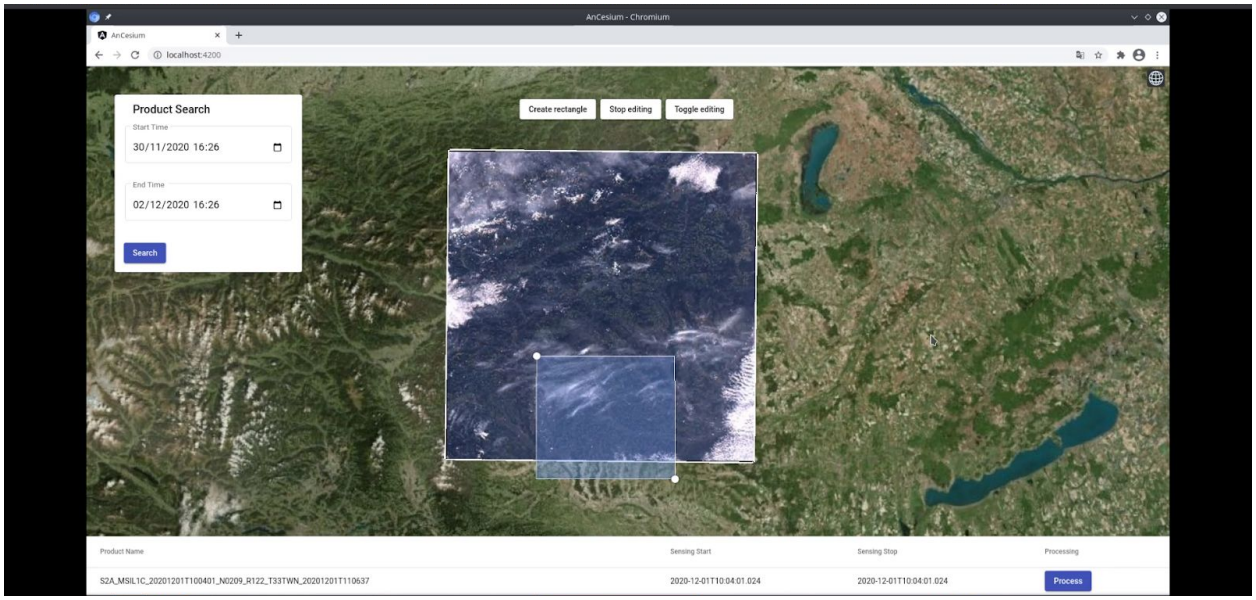
3.2.4. Service Endpoint

The initial web page will ask for the user credentials, the EGI checking will be used for such purpose. After that, a viewer screen (figure 7) will be shown to the user in order to enter the selection criteria for the data imagery search. It is possible to select an AOI over the globe and a period of time to perform the search of products over the specified criteria.



Ffigure 7. Viewer for AOI selection

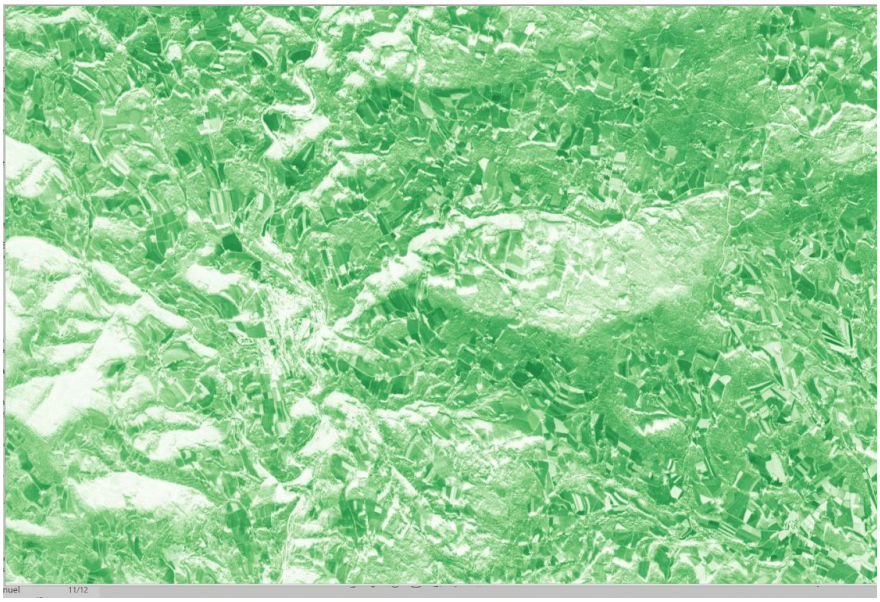
For the presented example, a direct search is performed to the ESA's data hub of Sentinel 2 imagery and the results of the search are shown in figure 8.



Ffigure 8. AOI selection and consult

The user can select the scene or desired scenes in order to send them to the processor. This is a simple example but the system is thought to be used for processing more advanced and complex processing of valued added products. The bottom of the screen shows the list of selected scenes and the button to command the processing.

Finally the user can retrieve the image locally and display it with the final result, as shown in figure 9.



Ffigure 9 NDVI processing result

3.2.5. Demonstration Video

The vídeo for G-CORE shows briefly a SaaS case to be offered in the EOSC market place for EO processing using G-CORE. The video is divided into two parts. The first shows a brief introduction of the functionalities available in G-CORE and the second is the video itself. The video shows the selection of an AOI area in order to select a scene to command the processing of a simple NDVI over that scene. After processing the scene the user displays and downloads the final product as a result of the processing.
The video can be found at the following link: https://youtu.be/xI9EJgq8m1Q

3.3. SAPS

3.3.1. Description

SAPS (SEB Automated Processing Service) is a service to estimate Evapotranspiration (ET) and other environmental data that can be applied, for example, on water management and the analysis of the evolution of forest masses and crops. SAPS allows the integration of Energy Balance algorithms (e.g. Surface Energy Balance Algorithm for Land (SEBAL) and Simplified Surface Energy Balance (SSEB)) to compute the estimations that are of special interest for researchers in Agriculture Engineering and Environment. These algorithms can be used to increase the knowledge on the impact of human and environmental actions on vegetation, leading to better forest management and analysis of risks.

3.3.2. Architecture

Figure 10 shows the architecture of SAPS. This architecture is automatically deployed, configured and managed by EC3. All the SAPS components run on a K8s cluster, so the location of each component depends on the K8s scheduler. The only component that needs to run in the front machine of the cluster is the Dashboard, so it can be exposed using the public IP of the front to the users.

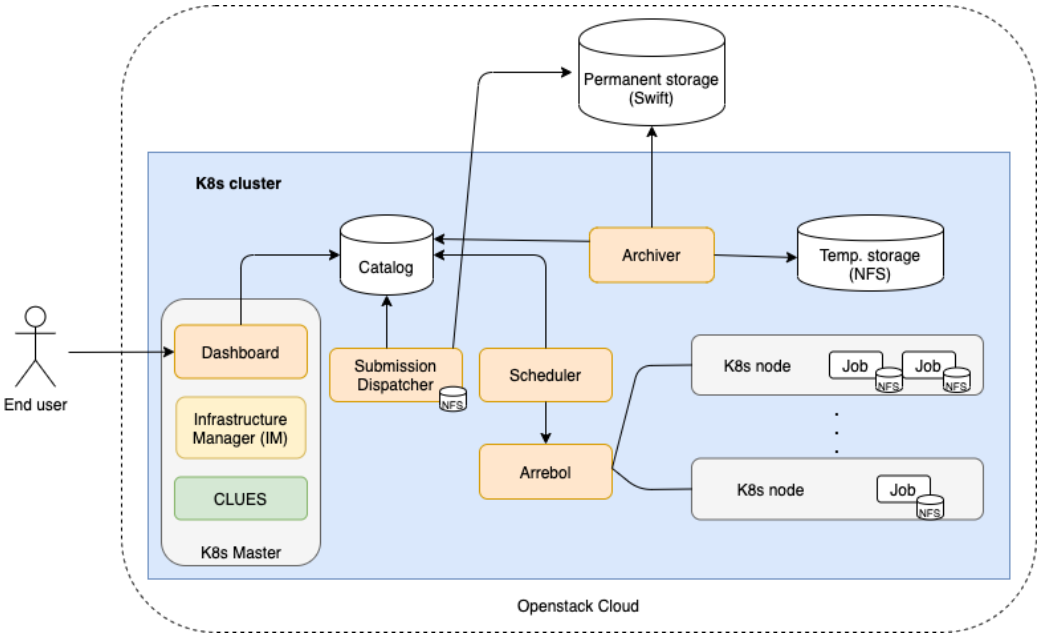


Figure 10 - Architecture of SAPS deployed on a K8s cluster by EC3.

As shown in figure 10, the user interacts with the system through the Dashboard, a web-based GUI that serves as a front-end to the Submission Dispatcher component. Through the Dashboard, the user, after successfully logging in, can specify the region, the period that he/she wants to process, as well as the particular Energy Balance algorithm that should be used. The execution consists of a three-stage workflow: input download, input preprocessing, and algorithm execution. With this data, the Dashboard

creates the processing requests and submits them sequentially to the Submission Dispatcher. Each request generated corresponds to the processing of a single scene. The Submission Dispatcher creates a task associated with the request in the Service Catalog database (PostgreSQL). This element works as a communication channel between all SAPS components. Tasks have a state associated with them that is used to indicate which component should act next in the processing of the task.

The Scheduler component is in charge of orchestrating the created tasks through various states until they finish. It uses Arrebol to create and launch the tasks on the K8s cluster as Kubernetes Jobs. A Job downloads the appropriate Docker image from Docker Hub and starts its execution. Input and output files are stored on a Temporary Storage NFS that is accessible to all Jobs running at the cluster. Arrebol monitors all active Jobs to find out the status of the executions, and updates the state of each task in the Service Catalog, accordingly. The Archiver component collects the data and metadata generated by tasks whose processing has either successfully finished or failed. The associated data and metadata are copied from the NFS Temporary Storage, using an FTP service, to the Permanent Storage, which uses the Openstack Swift distributed storage system, where they are made securely and reliably available to the users.

Through the Dashboard, the user can also have access to the output generated by completed requests. The interface to access the output data uses a world map. A heat-map, segmented based on the standard tiles used by the Landsat family of satellites, is superimposed to the world map. The heat-map gives an idea of the number of scenes for each Landsat tile that have already been processed.

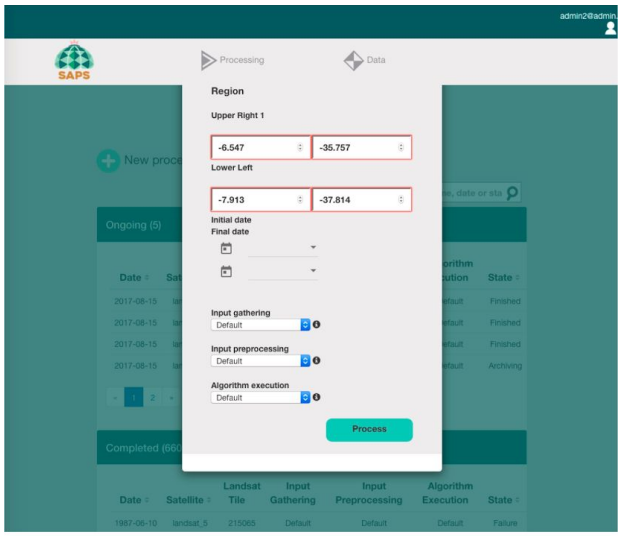
3.3.3. EOSC Services

In the context of EOSC-SYNERGY, SAPS is being integrated with several services offered by EOSC. This integration will facilitate European scientists to exploit the evapotranspiration estimation services from remote sensing imagery. Currently, the service relies on the next EOSC Services:

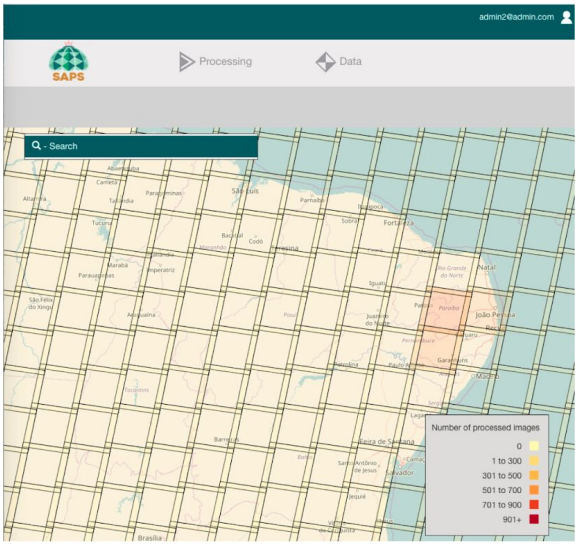
- **EC3 and IM** tools: both are services used by SAPS to deploy and configure a Kubernetes cluster automatically with SAPS running on it. Also, EC3 is used to manage the elasticity of the K8s cluster automatically. These tools facilitate the deployment and management of SAPS service.
- **EOSC computing resources**: through EC3 and IM, the SAPS service is deployed on top of a virtual elastic K8s cluster, that may rely on EOSC federated cloud computing resources or in on-premises solutions like Openstack.
- **EGI Check-in**: through EC3 portal. To deploy a cluster with SAPS, we use the EC3 portal of EOSC-SYNERGY, which is already integrated with EGI Check-in. So, to access a SAPS cluster, you should identify yourself with EGI Check-in. We will also consider integrating EGI Check-in directly on the SAPS dashboard in the next year of the project, for a fixed production endpoint of SAPS.

3.3.4. Service Endpoint

The SAPS dashboard is designed to facilitate the deployment and management of Landsat analysis tasks. Figure 11 shows the appearance of it for (a) submission of a new processing request and (b) access to the output data.



(a) Submission of new processing requests.



(b) Access to output data.

Figure 11 - Snapshot of the SAPS interface.

To access the SAPS Dashboard, a user is requested, as shown in figure 11. Internally, this is managed by local authorisation tokens. This solution is limited to the application, and we plan to study the viability of integrating EGI Check-in.

EOSC-SYNERGY does not provide an endpoint of the SAPS service. Instead, you can deploy your own instance easily through the EC3 EOSC-SYNERGY portal (https://servproject.i3m.upv.es/ec3-synergy/index.php), selecting as LRMS 'Kubernetes' and as Software package 'SAPS', or you can directly use the EC3 recipe and YAML files from the https://github.com/amcaar/saps-docker GitHub repository to deploy your own instance and properly configure the access to Openstack Swift storage solution. However, we also plan to offer a production instance of SAPS ready for non-advanced users, that will be available in the next months below the SAPS VO (*saps-vo.i3m.upv.es*).

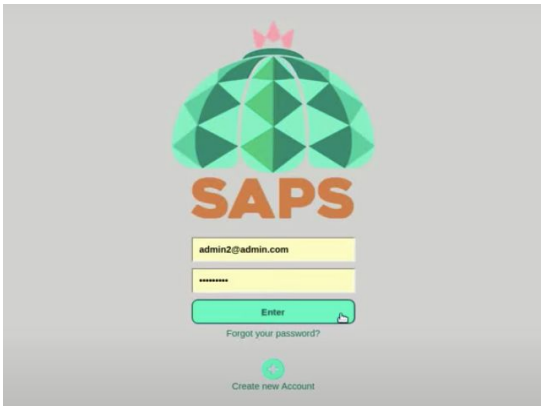


Figure 12 - Login screen of SAPS Dashboard.

3.3.5. Demonstration Video

We have prepared a demonstration video where we not only show SAPS in action, but its integration with some of the EOSC services (EC3 and IM) and the developments we have performed during the first year of the project (mainly the integration with Kubernetes). The link to the demo is:

https://www.youtube.com/watch?v=mM6xJJRS3Cs&t=17s.

The demo is mainly divided in three parts. The first part of the video shows the deployment of the SAPS application on top of an elastic Kubernetes cluster by terms of EC3. The video shows the command needed to deploy the cluster by using EC3 CLI and how to connect to it. Once inside the cluster, in the video we show how the SAPS microservices are deployed in Kubernetes and wait for an initial working node to run. This action is automatically done by CLUES, the elasticity manager of the cluster. On the second part of the video, we access SAPS Dashboard and show a bit the graphical interface it offers to create and monitor the status of the tasks. We also explain the required parameters SAPS asks the user to create a new landsat workflow analysis. Finally, the third part of the video shows an example execution created in the SAPS dashboard, and how the elastic Kubernetes cluster adapts automatically its size to cope with the 62 tasks that compose the workflow. The last part of the video shows the graphical interface that SAPS offers to access the output of the previously executed workflows, that is based on a world heat map.

3.4. SCIPION

3.4.1. Description

Scipion is an application framework developed by the Instruct Image Processing Center(I2PC) in Madrid to help the Structural Biology community to process CryoEM data. Scipion is developed as a plugin-based workflow management system that integrates many important software packages available in the field. Scipion is a desktop application and the user interacts with it through a GUI developed in python and Java. ScipionCloud, previously offered as an EGI AppDB image, has been converted to a web service that can be used by Instruct users to deploy a cluster in EOSC cloud resources to keep processing data acquired at a Cryo electron microscopy facility. This cluster has Scipion installed as well as most of the plugins and external packages used by the community, all ready to be used. In this first prototype only a single Scipion node is deployed but the next version will deploy a slurm elastic cluster using one of the core EGI services, EC3. Worker nodes will only be created when the workload requires it, ensuring an optimum usage of cloud resources.

3.4.2. Architecture

The architecture of SCIPION service is shown in figure 13. User acquires images at a microscope facility where some automatic preprocessing is done using Scipion. Raw data and projects are stored locally. Later, a user can access the IM dashboard portal using her ARIA credentials on the EGI check-in service and deploy a cluster with all software ready to be used.

The cluster comprises a front-end node with Scipion, a shared storage containing data and project (movies raw data might be skipped due to the size) and the required components to send jobs to worker nodes through SLURM batch system. In this first prototype slurm and EC3 are not yet fully integrated and a single node is deployed. Users will interact with the front-end server via noVNC from a web browser.

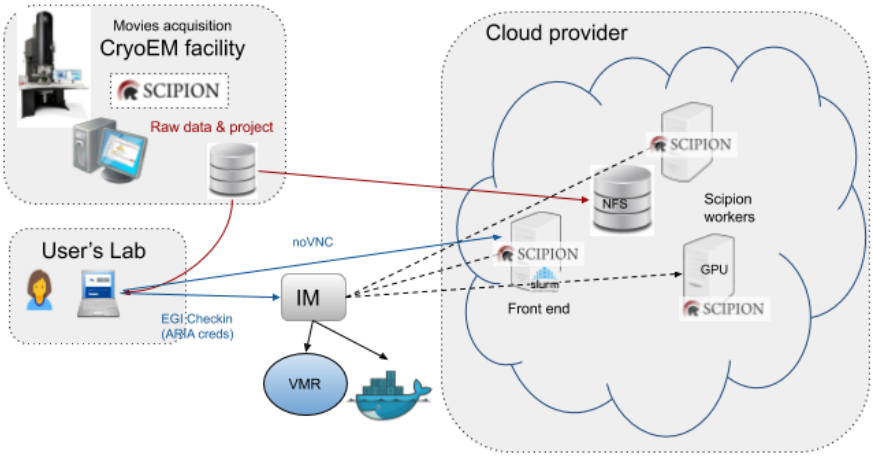


Figure 13. Architecture of first prototype for the SCIPION Service (only a single node but slurm cluster soon integrated).

3.4.3. EOSC Services

SCIPION service has moved from a static image that needed to be deployed directly on a specific site to a fully web service that is easy to use thanks to the integration of some of the EOSC core services. Now INSTRUCT users can access the service at the IM dashboard using her ARIA credentials and deploy a cluster on EOSC cloud resources by clicking on a button and providing some input through a guided wizard. All this is accomplished by using the following services:

- **Authentication**: The portal to deploy Scipion service (currently the IM dashboard) integrates the **EGI Check-in** service that allows INSTRUCT users to keep using her ARIA credentials.
- **Elastic cluster deployment:** Currently **IM** service is used to deploy a single Scipion node but the next version will use **EC3** and **SLURM** to deploy an elastic cluster on EOSC cloud resources or public clouds such as AWS EC2. Scipion can send jobs through SLURM which is totally integrated with EC3.
- **EOSC cloud resources**: The cluster might be deployed on EOSC federated cloud if user credentials permit it.

3.3.4. Service Endpoint

Scipion service is accessed through the IM dashboard login using EGI Checkin and ARIA credentials. Scipion is available in the IM collection of application templates as an icon to guide users through creation of the node (cluster) as shown in figure 14.

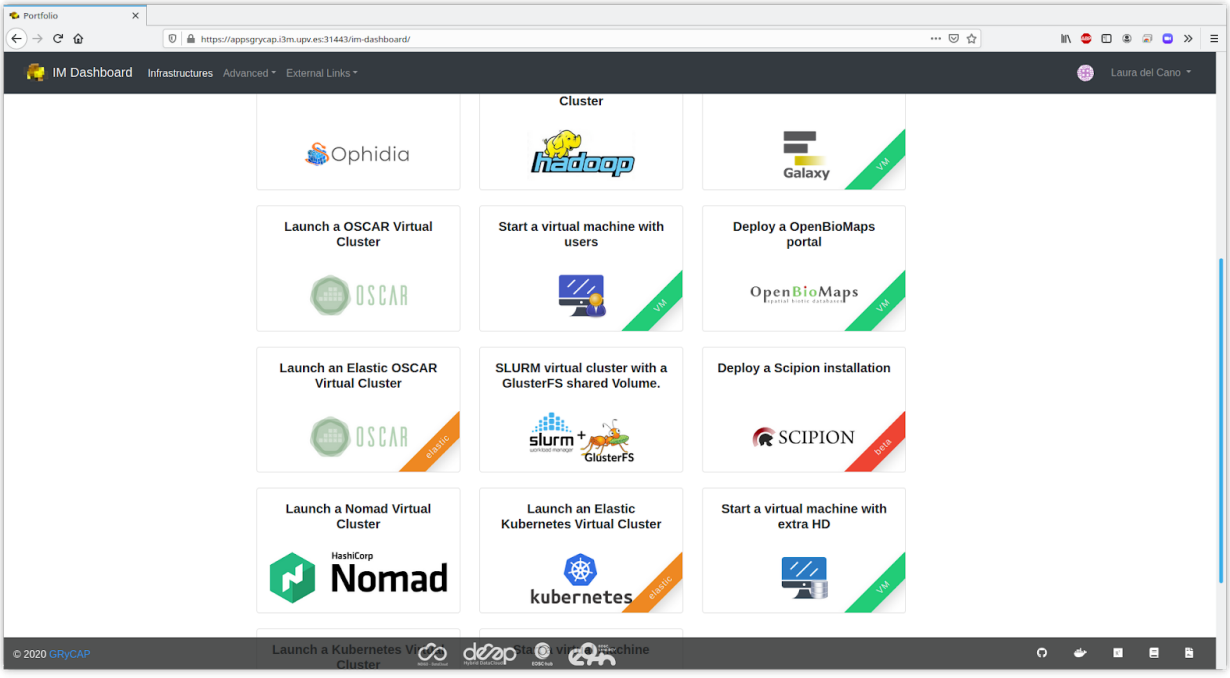


figure 14. Scipion available from the IM collection of application templates.

In this first version users only need to specify the VNC password to access the remote desktop in the front-end server but a field already present will allow uploading a Scipion project to continue processing in the cloud (see figure 15).

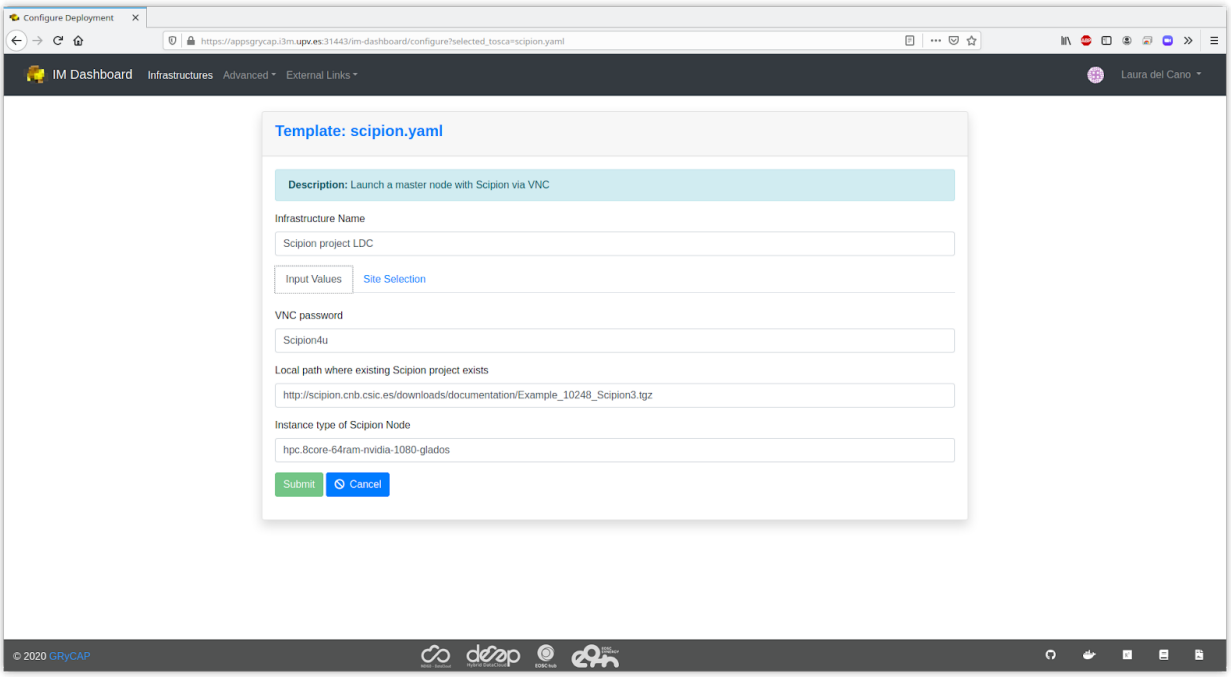


Figure 15. Scipion connection details.

3.4.5. Demonstration Video

A video has been prepared to show the current status of the service. It can be found in the following link:

https://youtu.be/0fiz23TRCpg

It demonstrates how to access the service using EGI Checkin and deploy a server in the cloud. Then the user can process her CryoEM data on a full Scipion installation in cloud resources, including GPU, which is a must for CryoEM processing nowadays.

3.5. OpenEBench

3.5.1. Description

OpenEBench (https://openebench.bsc.es) is the ELIXIR benchmarking and technical monitoring platform for bioinformatics tools, web servers and workflows. The development of OpenEBench is led by the Barcelona Supercomputing Center (BSC) in collaboration with partners across different European projects and Life Sciences communities.

OpenEBench as platform has the overall objectives:

- Provide guidance and infrastructure support for community-led scientific benchmarking efforts.
- Provide an observatory for software quality based on the automated monitoring of FAIR for research software metrics and indicators.
- Work towards the sustainability of the platform by adopting, integrating and promoting principles on Open Software, Open Data and Open Science.
- Adopt community-led standards, protocols and/or including the Global Alliance for Genomics and Health (GA4GH), ELIXIR and the European Open Science Cloud (EOSC).

Building on those objectives, OpenEBench can engage with different end-user profiles across the Life Sciences communities and beyond.

- **Developers**, who have a reference place to identify current challenges and relevant data sets for developing new algorithms and/or measure the impact of new developments. OpenEBench offers the possibility to developers to compare the scientific performance of their solutions with others from the community. Ultimately, it helps to improve their methods and disseminate their results thanks to publications and results spreading.
- OpenEBench assists **Communities** in the organization of their scientific benchmarking activities and the identification of new trends in their concrete area by providing examples of assessment metrics already in use in other communities, contributing to results dissemination and establishing good practices.
- **Researchers** mainly benefit from getting guidance about choosing the best resource for their research needs and be aware of the latest advancements in the area by getting information from trusted experts and staying up to date with new developments.
- **Funders** are able to maximize impact from projects which include the development of new software resources and/or improve the existing ones.

3.5.2. Architecture

As described in figure 16, OpenEBench scientific benchmarking architecture has three different levels that allow communities at different maturity stages to make use of the platform.

- Level 1 is used for the long-term storage of benchmarking events and challenges aiming at reproducibility and provenance. Level 1 makes use of the OpenEBench data model (https://github.com/inab/benchmarking-data-model), which allows organizing any relevant data

used and/or generated by community-led scientific benchmarking activities. Data is bundled and deposited in services provided by facilities like B2SHARE from EUDAT, where they receive a DOI. This enables full data provenance and reproducibility for everyone involved.

- Level 2 allows the community to use benchmarking workflows to assess participants' performance. Those workflows compute one or more evaluation metrics given one or more reference datasets. Workflows for level 2 are organized using software container technologies (e.g. Docker or Singularity), and computational workflows managers like Nextflow. This choice facilitates the deployment and use of level 2 workflows across any computational installation compatible with such technologies.
- Level 3 goes further by getting workflows specifications from participants, and then evaluating them in terms of technical and scientific performance. At this level, the whole benchmarking experiment is performed at OpenEBench; first, the predictions are made using the software provided by the participants; then, those predictions are evaluated with the benchmarking workflows; and, finally, the results are stored and visualized in the web server.

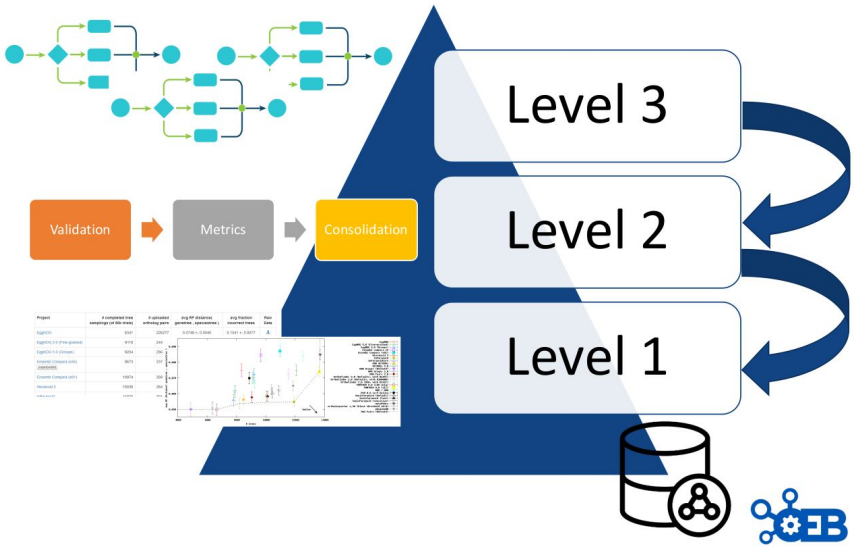


Figure 16. Conceptual diagram of support levels of OpenEBench.

Importantly, each level makes use of the architecture defined in the previous level e.g. participants' data generated by workflows at Level 3 are evaluated using the metrics and reference datasets in Level 2, and the resulting data is stored following the data model in Level 1 for private and/or public consumption.

3.5.3. EOSC Services

OpenEBench already uses ELIXIR AAI, which is intended to evolve together with other services e.g. GEANT; as Life Sciences AAI in the context of the cluster project EOSC Life. OpenEBench has started to incorporate some of EOSC Life services, specifically, WorkflowHub, as a mechanism to facilitate the provenance of the workflows used in the platform as well as a mechanism to monitor the availability and deployability of workflows used in OpenEBench within the community-led scientific benchmarking activities.

OpenEBench is integrating specific services from the EOSC Portal. Specifically, OpenEBench is integrating EUDAT services for the long-term availability of benchmarking results. To make this possible, EUDAT has created an OpenEBench Community, which will be used to associate any datasets from community-led scientific benchmarking activities. Using EUDAT allows us to assign a unique identifier, e.g. DOI, for those datasets contributed by members of communities at OpenEBench when publishing their results. In this way, it will be possible to reproduce at any time specific published benchmarking results by anyone interested. This integration requires the advanced management of users authorization as data should be deposited on behalf of their original owner rather than using OpenEBench identities.

It is expected that OpenEBench will become part of the EOSC portal portfolio by exposing and deploying the benchmarked analytical workflows as well as extending its capacity through best practices and additional services. As impact, we expect Life Science researchers will have semantically annotated, up-to-date collections of analytical workflows, which can be deployed across heterogeneous systems, organized by scientific communities around specific topics. As it is already happening, OpenEBench is contributing to organize emergent communities around scientific benchmarking activities by providing best practices and success stories of other communities.

3.3.4. Service Endpoint

This section includes some snapshots of the interface including the access procedure. Figure 17 (right) shows the result of a benchmark comparison, figure 17 (left) shows the access through Life Sciences AAI and figure 18 shows the user's workspace.

The service is available at https://openebench.bsc.es/. The Virtual Research Environment (VRE) is accessible at https://openebench.bsc.es/vre/tools/QFO_6/input.php?op=0. In this VRE, the users can upload their own data and applications for the execution of the benchmarks. In the general thematic service, any user can browse the information related to the benchmarks registered in the platform.

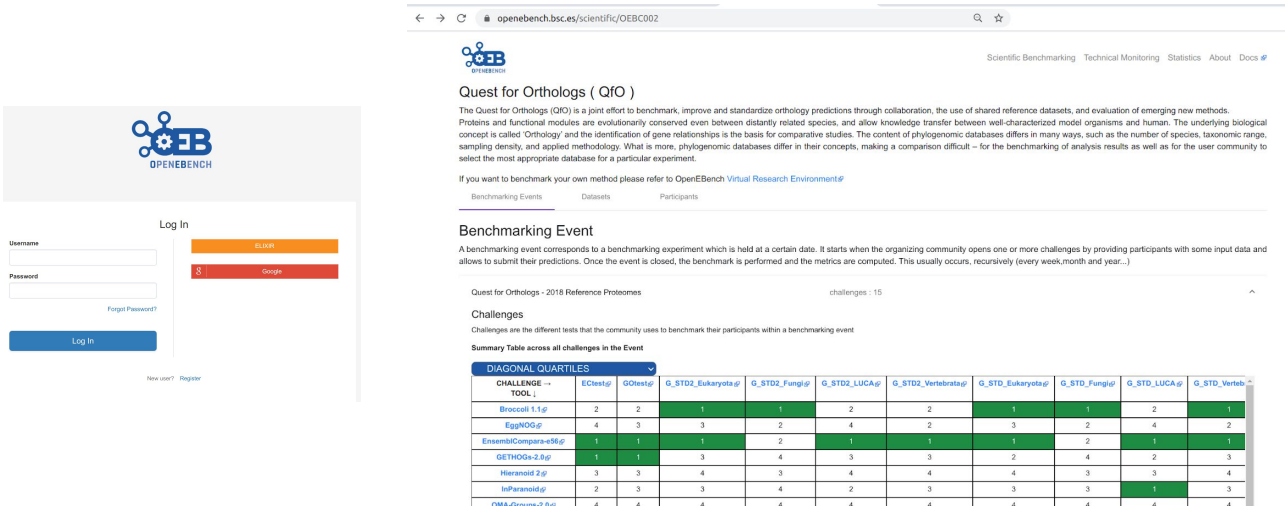


Figure 17. Screenshots of OpenEBench with the Sciences AAI access (left) and with the comparison of different methods for a specific benchmark using OpenEBench (right).

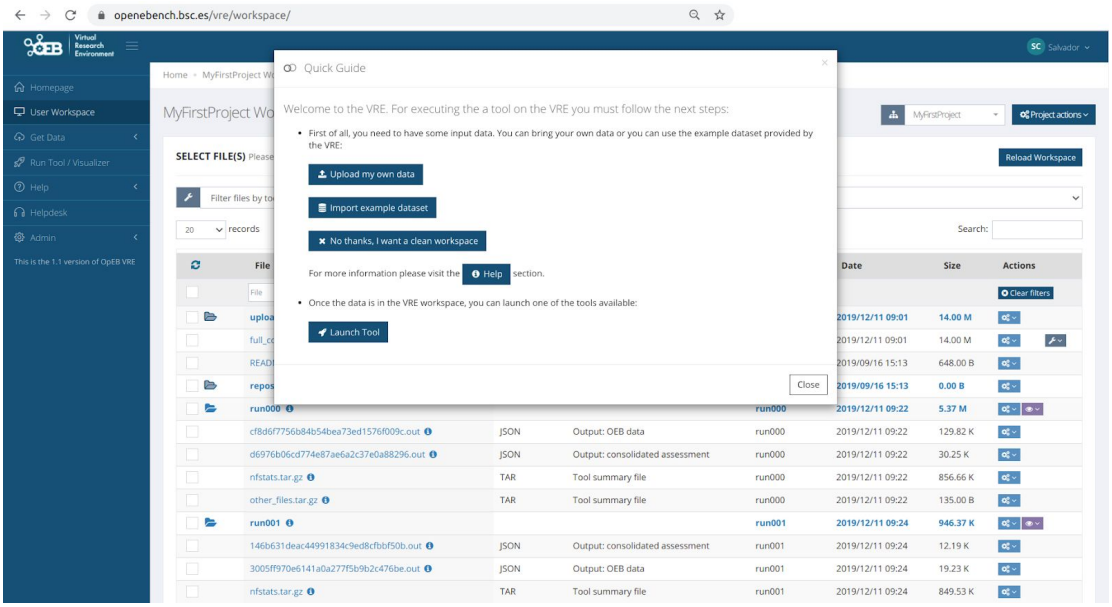


Figure 18. Screenshot of the workspace in OpenEBench.

3.5.5. Demonstration Video

In this video we demonstrate the use of EOSC Synergy services within OpenEBench, specifically, the use of B2SHARE, a EUDAT infrastructure meant for long-term storage and sharing of research data.

The video is divided into two main parts. The first part presents an overview of the platform and explains its main components and architecture, while the second one describes how to upload one of the files used in a benchmarking workflow from the Virtual Research Environment to B2SHARE. The video can be found at the following link: https://youtu.be/YNUziurPBIc

3.6. LAGO

3.6.1. Description

The LAGO (Latin American Giant Observatory) Project is an extended astroparticle observatory at a global scale. It is mainly oriented to basic research on three branches of astroparticle physics: the extreme universe, space weather phenomena, and atmospheric radiation at ground level. Parallelly, these are the needed components of other research on high energy physics, weather forecasting, life sciences, aerospatial security or computer science.

The LAGO detection network consists of single or small arrays of self-designed water-Cherenkov detectors (WCDs). These particle detectors are spanned over different sites located at significantly different latitudes (currently from Mexico up to the Antarctic region) and different altitudes (from sea level up to more than 5000 meters over sea level), covering a huge range of geomagnetic rigidity cut-offs and atmospheric absorption/reaction levels. The measurements collected from these detectors are posteriorly processed and analysed. Moreover, scientists continuously generate simulated data for arbitrary locations and weather conditions.

On the other hand, the LAGO Project is operated by the LAGO Collaboration, a non-centralized and distributed collaborative network of more than 100 scientists from almost 30 institutions in 11 countries. Additionally, several universities have incorporated LAGO studies into their curricula. Their students, especially the ones belonging to physics, electronics and computing areas, also contribute to the development of LAGO technologies.

To manage this heterogeneity and take advantage of the aforementioned contributors, the LAGO Thematic Service will progressively incorporate the continuous generation of data (measurements, processing and simulations) and code into standardised mechanisms that follow the FAIR principles. This is so to guarantee the long-term curation and re-use of data as well as the dissemination or reproducibility by other communities.

Therefore, the final purpose of the LAGO Thematic Service is to enable the universal profit and contribution of this research, within and outside LAGO Collaboration, through a sustainable Virtual Observatory and standardised computational model.

3.6.2. Architecture

To introduce the architecture of the LAGO Thematic Service, readers should understand first some basic considerations about the kind of data managed and the target community.

There are two main kinds of data managed by LAGO Collaboration. The first is related to real measurements (L), and the second is to simulations (S). Thus, the measured data (raw) is pipelined for correction, obtaining the following data sub-types that corresponds with their quality level:

- **L0. Raw data**. Measurements of Water-Cherenkov detectors (WCDs).
- **L1. Preliminary data**. Low resolution but the atmospheric pressure is corrected.

- **L2: Quality for Astrophysics**. Ensures data quality to be used by experts from the astrophysics Community: fixed scalers by atmospheric parameters and the efficiency of the detector.
- **L3. Quality for the public**. Ensures high quality to be used by researchers from other subjects or the general public: the histograms are also corrected.

On the other hand, users can perform their simulations of rains, generating two sub-types of data-sets:

- **S0. Plain simulations**. Plain simulated data (CORSIKA outputs managed by ARTI).
- **S1. Analysed simulations**. ARTI outputs.

There are four main premises of the collaborators:

- Officially, they are grouped into **autonomous research units** within work packages, every unit with specific responsibilities. As examples, every detector has an operator unit, every software piece has a manager, etc. External staff should be allowed or removed by every research group for eventual contributions.
- Most are **researchers** in astrophysics and HEP **with a background in computing skills**: they are accustomed to profiting from HPC facilities and/or use control version systems such as Git.
- Although each contributor is focused on simulations, on processing or curating measurements, **they produce results of interest for any other member or external actors: whole data or code generated should be registered**, shared and published after an embargo period.
- Some **institutions** support the project **with computational resources**, such as clusters and related storage, but they generally are non-exclusive and provide a limited environment.

The design should take in mind the aforementioned tasks as well as the thematic service is focused on providing a standardised way to curate and reuse measurements, analysis and simulations. To achieve this task, the architecture follows the basic **design recommended by EGI/EOSC for cloud computing**: **core** intelligence **packed in Docker images**, being able to automatically check, store and publish their **results in DataHub**, with enough metadata to be **referenced by PiDs** (provided by **B2HANDLE**) and used by official harvesters (**B2FIND**), which will act **as virtual observatories**. As the whole computation is self-contained in the image, the **production** can be easily performed on cloud resources **deployed by services** such as **EC2/IM**, or even manually in private clusters.

The researcher's point of view of the architecture is shown in Figure 17. He can access to whole data and additional computational resources if he first logs into the **LAGO Virtual Organisation** (VO), which is managed through **eduTEAMs**. Then, he can create or transform old data if he uses the official Docker images for these tasks. Additionally, he could run these Dockers on any cluster outside EOSC or inspect inside them, easing the bug resolution. The only condition is to use the official Docker images and to register results in DataHub. This is so to ensure the correct generation of metadata for every dataset, which will posteriorly enable the publication, searches and data-mining. Moreover, WCDs continuously provide raw data, which also are under a certain user/group's responsibility registered in the VO. These imported data also should be checked and completed with metadata, which is performed by its specific Docker instance.

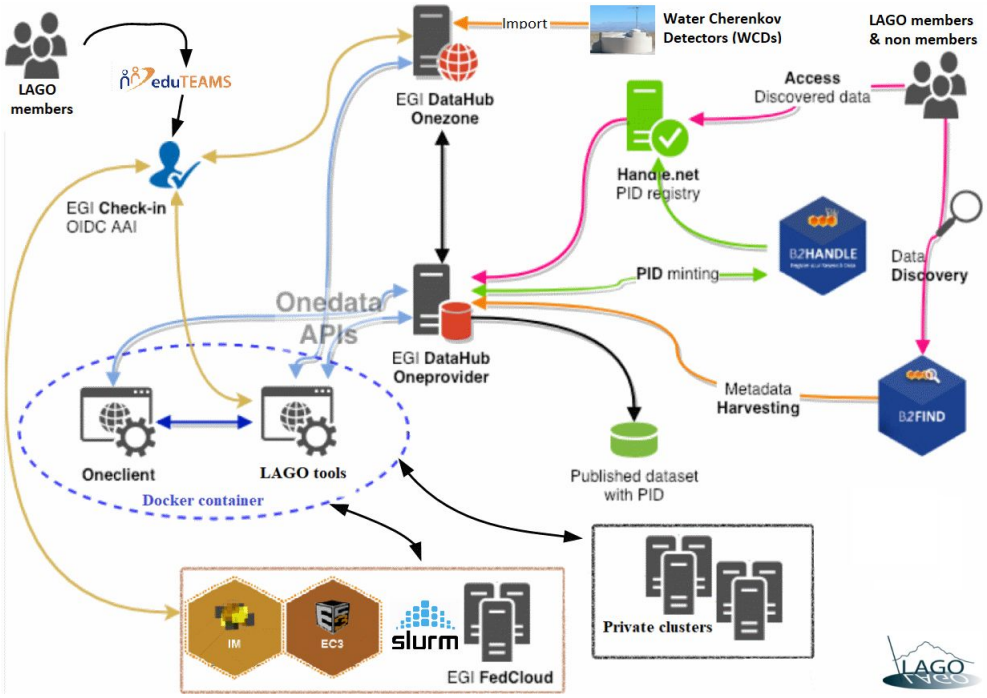


Figure 17. The architecture of the LAGO thematic service.

Therefore, besides the integration of these services and the creation of a new Virtual Organisation, the core contributions of the LAGO thematic service are focused on the definition of metadata and its generation by LAGO tools deployed in Docker images.

Note that only the **L0, S0 and S1** types of data will be covered in EOSC-SYNERGY. Therefore only related metadata and LAGO tools will be migrated to Docker during the project, although the architecture will be maintained when adding the rest of computation in the future. Additionally, for this **prototype**, only S0 simulations are supported, being the process that consumes more computational resources and it implies to overcome more difficulties in its deployment. Its Docker instance, named **OnedataSim** (https://github.com/lagoproject/onedataSim), encapsulates ARTI and CORSIKA software, generates the data and metadata and stores them in DataHub. It is currently available for the whole LAGO community.

The schemas used and definitions have been published in the mandatory data management plan of LAGO (https://lagoproject.github.io/DMP/). The main characteristics are the following:

- Language syntax: **JSON-LD 1.1** (W3C). It is the simplest standard for linking metadata. Promoted by Google, currently, it is ousting more heavy and complex syntaxes such as RDF, Turtle, XML.
- Main vocabulary: **DCAT-AP2** (European Commission), which is a specific profile of DCAT2 (W3C), recommended for repositories, content aggregators or data consumers related to the public sector (government, research centres, funded projects).
- The **LAGO vocabulary**: described in this document. It is a re-profile of DCAT-AP2, extending the existent classes and adding properties needed for LAGO computation.

- Catalogues and datasets: Every **file** generated is considered **the minimum data-set** to be data-linked and processed, while **a collection** of related files is grouped in a **catalogue**, which should be referenced with a persistent and unique identifier (PiD). As the different LAGO activities generate only one data sub-type, **catalogues will only contain files belonging to one sub-type an activity**, with exception of checking or correction procedures.

3.6.3. EOSC Services

LAGO thematic has selected and it is integrating the following services listed in the EOSC marketplace:

- **EGI Check-in (through EduTeams Perun at GEANT)**: it is needed for accessing any EOSC service, in particular for obtaining a OneData token. However, managing the VO with Perun at GEANT was considered because of flexibility, certain independence from EU Framework projects and long-term support to Latin American users. Perun provides the needed flexibility allowing several roles and permissions over the data, such as conventional users (allowed seeing whole data, restricted write), research group chiefs (allowed enrolling their researchers by their own), robots, main administrators, etc. On the other hand, the sustainability of the VO is guaranteed by the support of RedClara (associated with GEANT), allowing extending users and resources beyond EOSC.
- **EGI DataHub**: OneData allows researchers several ways to access the data and metadata of their interest. Collaboration members can directly explore the directory tree at https://datahub.egi.eu or mount it on their PC's. Meanwhile, the general public will get published data through B2FIND. On the other hand, OneData eases storing results without modifying simulation/processing codes, as well as maintaining usable replicas around the world. Currently, DataHub is storing S0 simulations with metadata and L0 raw without metadata, taking up over one TB.
- **The EOSC Cloud services (IM and EC3):** simulations are arbitrarily performed by researchers running the dockers in EOSC Cloud services. To perform these tasks, they dynamically deploy individual virtual machines or batch clusters through IM or EC services. Although users can create any kind of cluster, only **Slurm** workload manager is supported for now, because it is commonly used by LAGO collaborators.
- **B2FIND and B2HANDLE:** currently under development, will be adopted in the coming months because we expect that the integration will be straightforward since we use standard metadata. In the case of B2FIND, we do not discard to additionally use other CKAN repositories in the future, such as the ones used in the EU Joint Research Centres and other government repositories, to completely benefit from the linked-metadata in JSON-LD + DCAT2-AP format.

3.6.4. Service Endpoint

As mentioned in subsection 3.6.2, collaborators in LAGO commonly use the command-line shell to run many scientific applications only available on Linux, which are needed for their research. Additionally, they like to inspect configurations, code and results for debugging. Moreover, they usually take advantage of remote HPC facilities. Thus, it is reasonable to offer a similar environment for them.

For these reasons, neither a customised submission web page nor portal (i.e. Galaxy) is built. However, the whole infrastructure is offered as the grid computing fashion: every user can deploy his cluster. Researchers only need a guide (https://lagoproject.github.io/DMP/howtos/) to enrol into the LAGO VO, to build a cluster in the cloud, to run standardised Docker instances in the batch system, and finally to check the results in DataHub. Additionally, every standardised Docker has a specific guide at their code repository, as it is the case of this prototype, onedataSim (https://github.com/lagoproject/onedataSim). Therefore, these are the specific endpoints for the LAGO Thematic Service, but real service endpoints are the EOSC services offered (IM:https://appsgrycap.i3m.upv.es:31443/im-dashboard; EC:https://servproject.i3m.upv.es/ec3-ltos/; B2FIND:http://b2find.eudat.eu; and DadaHub: https://datahub.egi.eu).

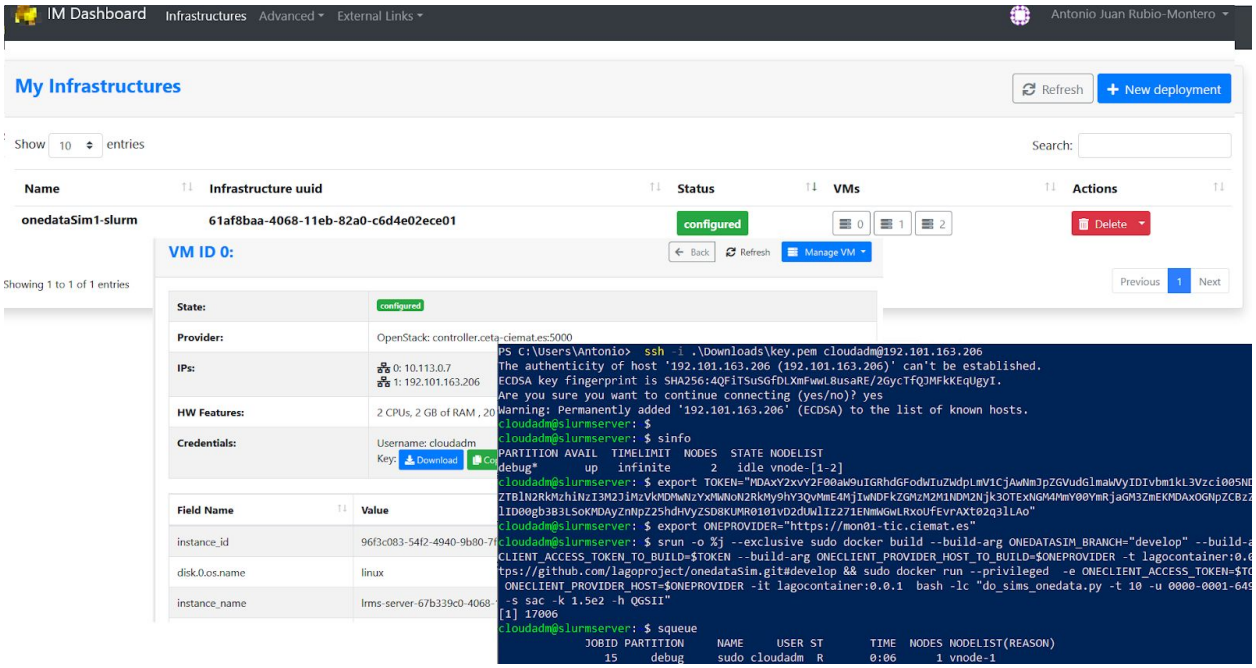


Figure 18. An example of running onedataSim in the Slurm deployed through the IM service.

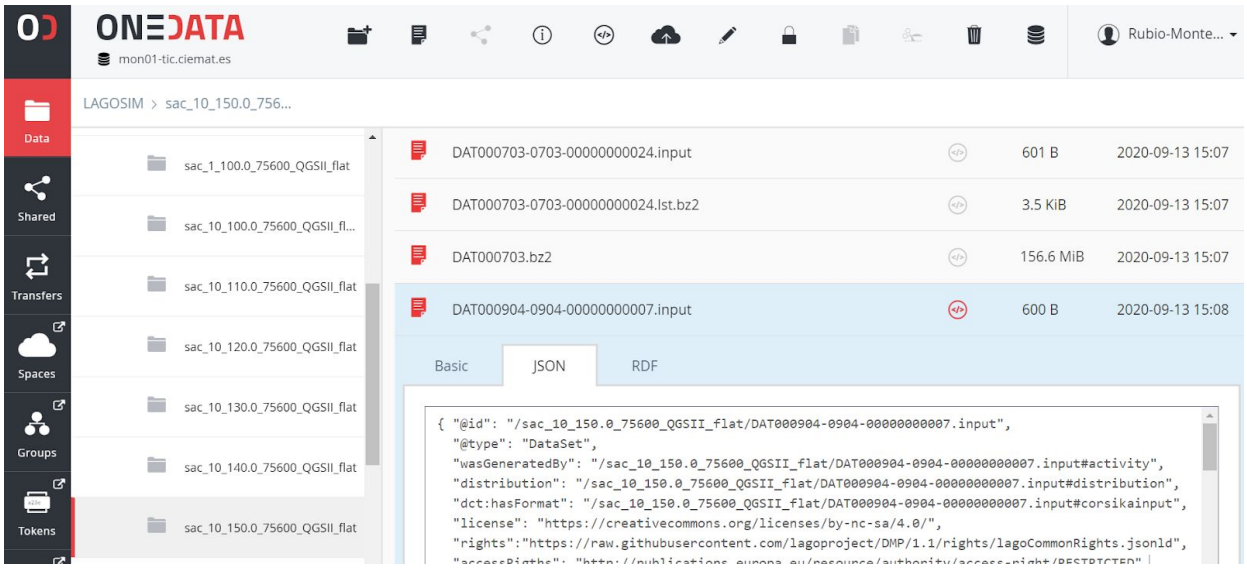


Figure 19. S0 metadata and data stored in DataHub by simulation shown in figure 18.

3.6.5. Demonstration Video

The demonstration video describes the LAGO project and the final purpose of the thematic service associated with the present deliverable. The video is available in https://youtu.be/LjP-fxv5_rQ

The video includes an explanation on how the EOSC-SYNERGY project is helping LAGO to incorporate the continuous generation of data (measurements, processing and simulations) and codes into standardised mechanisms that follow the FAIR principles. The video describes the different types of data generated by the virtual astroparticle observatory and the computational models and presents the chosen technology solutions and architecture of the management plan. Additionally, includes instructions on how to join the LAGO organization and how to access the data repositories. Finally, the video includes a tutorial on how to access the infrastructure manager and the description of the tools for ARTI simulation and analysis on OneData.

3.7. SDS-WAS

3.7.1. Description

SDS-WAS provides a set of services related to the mineral dust forecast. It collects numerical model outputs and observational data from a wide set of worldwide partners plus internally developed. A wide set of post-processed analysis and statistics are generated, and results in form of plots, tables or numerical (binary) data are disseminated to a variety of users (public institutions, researchers, etc ...). Finally, SDS-WAS also organizes training courses on dust related research, to give capabilities for generated products usages and other kinds of events (seminars, workshops) related to dust research. The aim is to give support to institutional entities (e.g. National Meteorological Agencies) to warn about possible dust events and to foster the study of dust-related phenomena into the academic and research communities. The EOSC infrastructure will give the possibility to increase the volume of data hosted and processed, reach a wider set of end-users, improve data FAIRness and robustness of the whole service infrastructure.

3.7.2. Architecture

As shown in figure 19, the service is configured as an High Availability Cluster of two duplicated instances. Each instance, one hosted in the Barcelona Supercomputing Center, and the other in AEMET (Spanish Meteorological Agency), runs in the front-end a Content Management System which delivers all web contents, and where all authenticated users have access to add/remove/modify contents. There are different levels of access (through a classic web log-in) according to roles of each user or group. Currently all products generated are disseminated through the web, maps, plots, calculated statistics and numerical data (models outputs).

Some of these products are pushed to different channels in background (WMO Global Telecommunication System: https://www.wmo.int/pages/prog/www/TEM/GTS/index_en.html and EUMETCAST: https://www.eumetsat.int/eumetcast).

In the backend a local storage hosts all data, received and generated, and a set of libraries and tools for post-processing and data analysis. Currently the service downloads data and produces its outputs (plots, numerical data, ...) on a daily basis with a set of cron-jobs. In-house models are run separately in two HPC clusters and the front-end accesses one of those outputs by default, and the other if the first one is not available.

On the other hand, the new refactored services is going to have:

1) An automatic and independent workflow which will manage the in-house model run in multi-HPC environment
2) A single sign-on system provided by the authentication service B2Access, which will be the entry point
3) All numerical data products will be available through the THREDDS data server integrated into B2SAFE storage service. The data server is specific to netCDF data format, which is the one used

in the SDS-WAS and one of the most popular for atmospheric science data. It exposes public APIs and provides aggregation features to data, independently of how files are physically organized into the file system underneath. It provides one entry point per data set and features, among others, to select dynamically spatial and temporal subsets. The B2Handle service integrated with B2SAFE will manage the Persistent IDentifiers generation and management.

4) A new interactive application connected directly to the B2SAFE repository which will replace (partially) static plots generated by nightly jobs. This application will, among other things:
    a) Show comparison view of forecast plots of different models selected dynamically by the user (figure 20)
    b) Show compared timeseries of selected model in a selected point (fig_sdswassdash2)
    c) Show evaluation plots of model outputs against observations
    d) Show verification scores tables of models against observations performances (BIAS, RMSE, etc ...)
    e) Show other relevant products (probabilistic forecast maps, warning advisory for specific countries, etc ...)

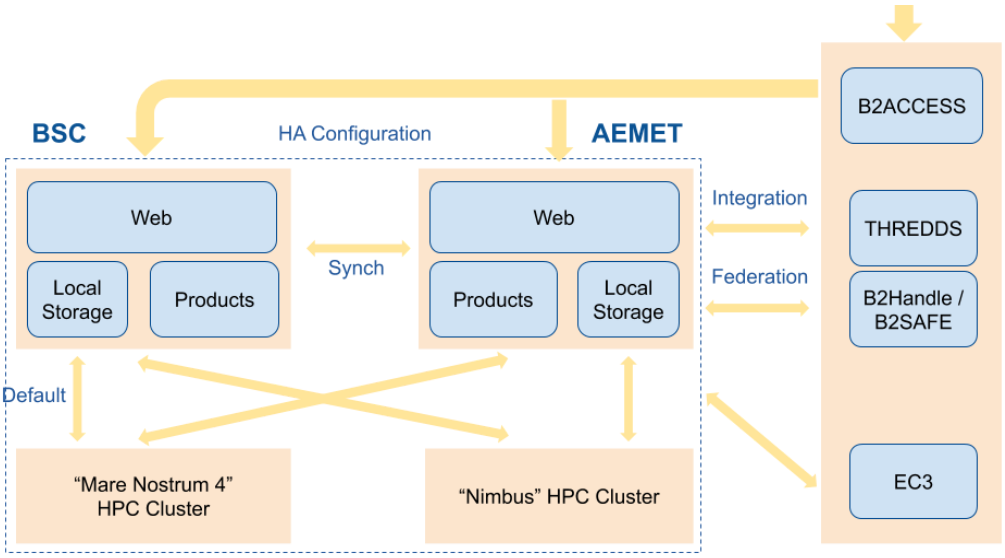


Figure 19. Architecture of the SDS-WAS thematic service.

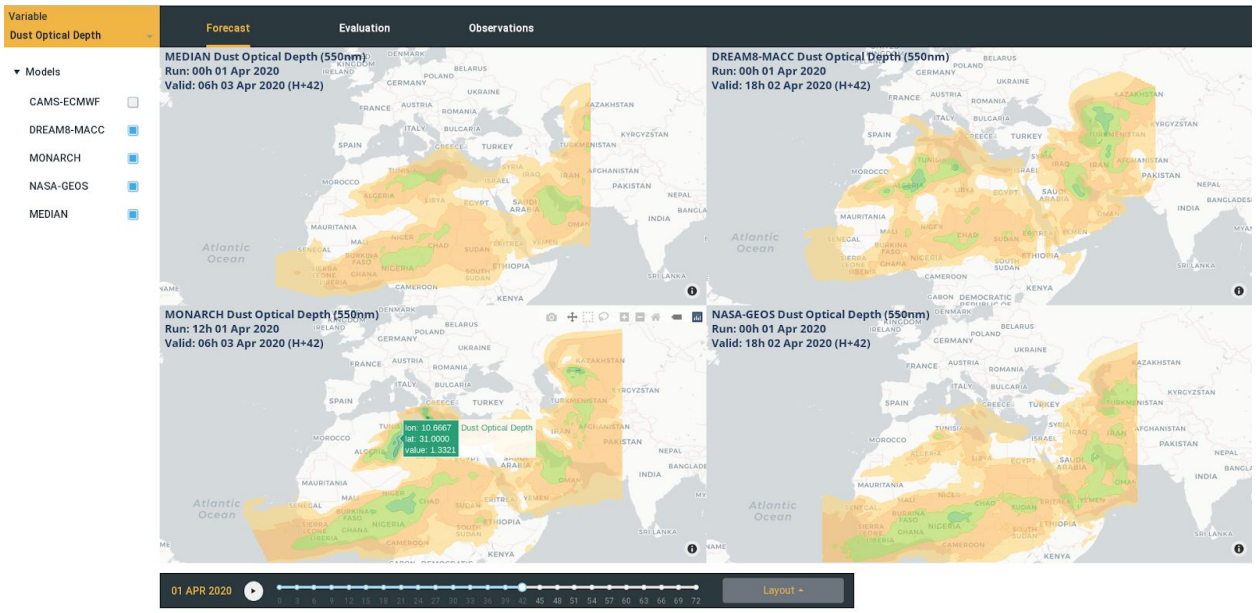


Figure 20. New interactive dashboard with models forecasts comparison plots.

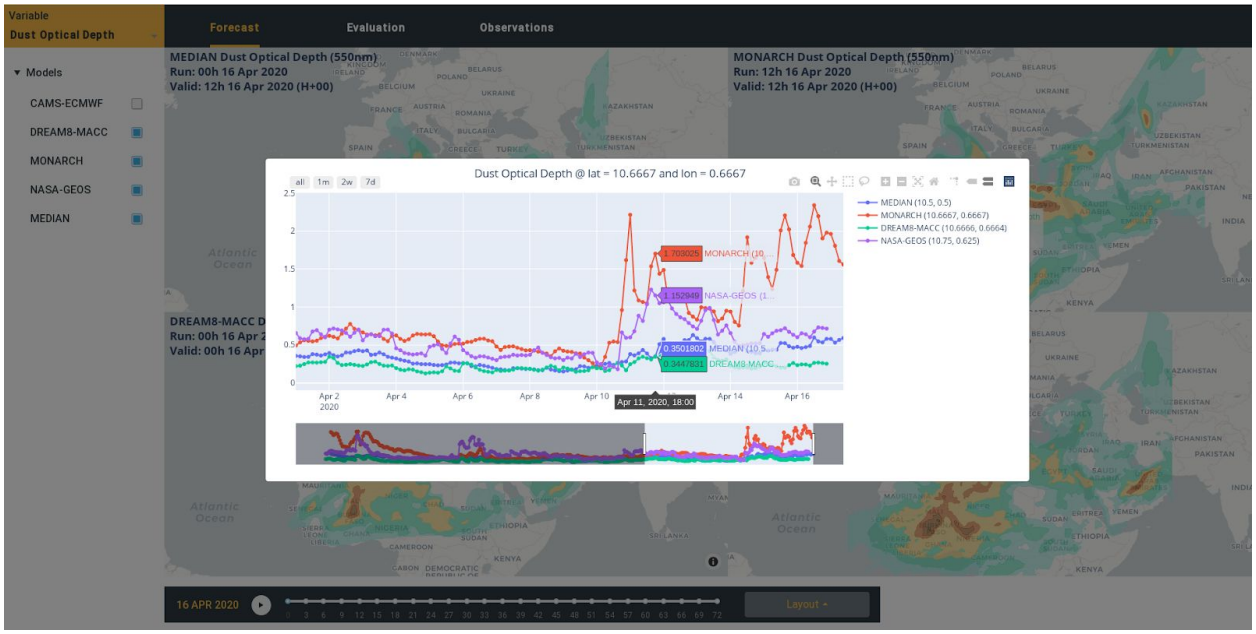


Figure 21. New interactive dashboard with models forecasts comparison time series.

3.7.3. EOSC Services

SDS-WAS thematic service has selected and it is integrating the following services listed in the EOSC marketplace:

- B2ACCESS will be used as a single sign-on platform to authenticate users into the services.

- B2Handle will be in charge of managing and generation of Persistent IDentifiers for data products delivered.
- B2SAFE runs as a storage service and will expose data products through the integration with a THREDDS (https://www.unidata.ucar.edu/software/tds/) data server.

3.7.4. Service Endpoint

The current interface of SDS-WAS is the http://sds-was.aemet.es website. All operational services are available through that end-point, such as:

- Forecast plots: https://sds-was.aemet.es/forecast-products/dust-forecasts/forecast-comparison
- Forecast evaluation: https://sds-was.aemet.es/forecast-products/forecast-evaluation/model-evaluation-metrics/model-evaluation-metrics-v3
- Files download: https://sds-was.aemet.es/forecast-products/dust-forecasts/files-download

The whole system is being refactored in collaboration with AEMET and in the framework of EOSC-SYNERGY project according to the architecture described in 3.7.2 paragraph and with the integration of EOSC services listed in 3.7.3 paragraph. The temporary end-point deployed for new refactored services is: http://dust02.bsc.es (can be down from time to time). It will show the web Content Management System and the interactive dashboard shown in figures 19 and 20.

3.7.5. Demonstration Video

The video follows this structure:

1. A first introduction to the current operational SDS-WAS service as it runs now, with:
    a. Models forecasts comparison plots
    b. Numerical model data storage and download
    c. Forecast evaluation against observations time series
2. A second part on the developments for the new service completely refactored with:
    a. Storage moved to B2SAFE service
    b. The THREDDS data server build on top of it to distribute numerical data dynamically aggregated
    c. The automatic workflow of the in-house numerical model MONARCH which runs in multi-HPC environment
    d. The interactive visualization application built on top of B2SAFE storage service and which provides all services related to plots, going to replace currently delivered as static png generated with nightly jobs

Link here:

https://dust.aemet.es/Members/francesco.benincasa-40bsc.es/eosc-synergy-d4-2-demo-video

3.8. UMSA

3.8.1. Description

UMSA is an untargeted mass-spectrometry analysis service from RECETOX (Research Centre for Toxic Compounds in the Environment at Masaryk University) in the Czech Republic. The service is evolving to a key component of the emerging EIRENE ESFRI. By means of the integration in EOSC, uniform access to data and computing resources are provided, scaling the service to the target European-wide user community. Typically, mass spectrometry is done in a targeted way to confirm or disprove the presence of a specific compound in a sample. On the contrary, we aim at processing data to correlating the whole spectra (ie. all the present compounds) with other data (social, medical, other sample analyses, etc.) to work with more complex hypotheses of environmental impacts on human health.

The data are unrecoverable, original samples cannot be re-acquired, therefore long-term data storage (even decades) is required, together with appropriate data curation. Tracking provenance of the secondary (derived) datasets (what was the exact process of generating them from the original source data), is fairly critical, as the results may differ dramatically with different settings. Similarly, the exact links between datasets and physical samples they originate from must be maintained.

The current release provides a Galaxy workflow based on re-factored tools originating from Emory university (apLCMS and xMSAnnotator), which detect peaks in the input spectra and matches those to metabolite and pathway databases. Extending the workflow with auto-tuning peak picking parameters (based on the original xMSAnalyzer tool) is in progress. The workflow continues with filtering false positives by predicting their chromatographic retention time (adapted Retip tool) based on computing chemical descriptors and machine learning with respect to a large database of known compounds (doi: 10.1038/s41467-019-13680-7). With respect to the previous release, the service was extended with a set of tools supporting gass chromatography MS (XCMS, RamclustR, MatchMS) and several tools to support conversions among various chemical identifier standards.

3.8.2. Architecture

The service is deployed as a virtual cluster using an Infrastructure Manager RADL recipe. The cluster consists of a single head node, running Galaxy frontend and the Slurm server, and an arbitrary number of Slurm worker nodes. Data is shared among the nodes over NFS.

The deployment of the head node registers a configurable dynamic DNS name (umsa.dyn.cerit-sc.cz currently) to point to its assigned IP address. The well known service endpoint (umsa.cerit-sc.cz) is expected to be an alias pointing to the dynamic one. In the next step a *Let's encrypt* certificate is acquired to allow smooth https connection. The RADL recipes also require simple provider-specific customization (cloud network names and base image identifiers in particular).

Authentication of the end users is managed by ELIXIR AAI via its corresponding Galaxy module. Configuration of Elixir AAI is the only manual step in the service deployment; the policy of introducing a new service to Elixir AAI requires human approval and exchange of secrets which cannot be automated

so far. In the current version, the service can be accessed *bona-fide* -- all users who pass ELIXIR AAI authentication are allowed. However, we plan a trivial registration procedure (using the ELIXIR tools) to restrict the access.

The payload of the service are several tools in Galaxy. In order to keep strict control on the complex software dependencies, we wrap all the software to Docker containers; Galaxy and Slurm are configured to execute them in this way only. The tools themselves are installed from standard Galaxy toolshed to follow common procedures.

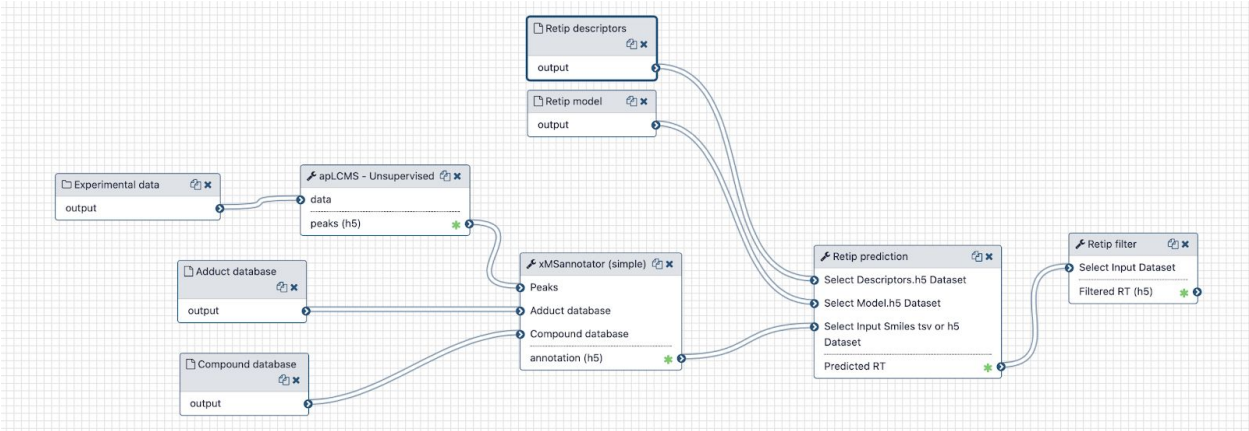


Figure 22. UMSA LC/MS Data processing Workflow.

The diagram in figure 22 shows the essential mass-spec data processing workflow. Besides the user input (.mzML spectra files) the workflow needs to query a metabolite database. In order to achieve reproducibility, timestamped snapshots of the online databases are used. They are downloaded and processed only from time to time, and the snapshots are available to all users. Most of the intermediate files passed between the tools are some kind of tabular data, using the HDF5 or Apache feather format.

3.8.3. EOSC Services

UMSA leverages the following EOSC services

- Infrastructure Manager (IM) and Elastic Compute Clusters in the Cloud (EC3): except from minor provider-specific customization the service is deployed with a generic RADL recipe submitted to IM, and IM takes care of its lifecycle.
- EOSC computing resources: the deployment relies on cloud computing resources; due to non-trivial computational demand of the workflow, "HPC" flavors of cloud virtual machines (with no CPU overprovisioning) are prefered. The production service runs at the CESNET/Masaryk University cloud site, we managed to deploy at CESGA successfully as well.
- EGI CheckIn: used to authenticate to IM as well as to deploy the VMs to the cloud sites. The service is currently using the "catch all" eosc-synergy.eu[1] virtual organization. A dedicated UMSA [2] VO was established recently and the service will migrate to it in early 2021.

[1] http://operations-portal.egi.eu/vo/view/voname/eosc-synergy.eu
[2] https://operations-portal.egi.eu/vo/view/voname/umsa.cerit-sc.cz

3.8.4. Service Endpoint

The principal service endpoint is https://umsa.cerit-sc.cz/. The user interface is standard Galaxy with minimalistic visual branding. Login with Elixir AAI credential is required as described above.

The individual tools are available in PeakPicking and Annotation sections, both exist in simple and advanced forms to address the needs of different user experience levels. The in-line documentation provides extensive description of the meaning of numerous input parameters, as well as appropriate references to web pages with further documentation of the tools, and the essential journal papers.

Figure 23 shows a typical input form of a simple workflow connecting the tools together. Figure 24 shows execution of this workflow in progress, generating its output files in Galaxy history.

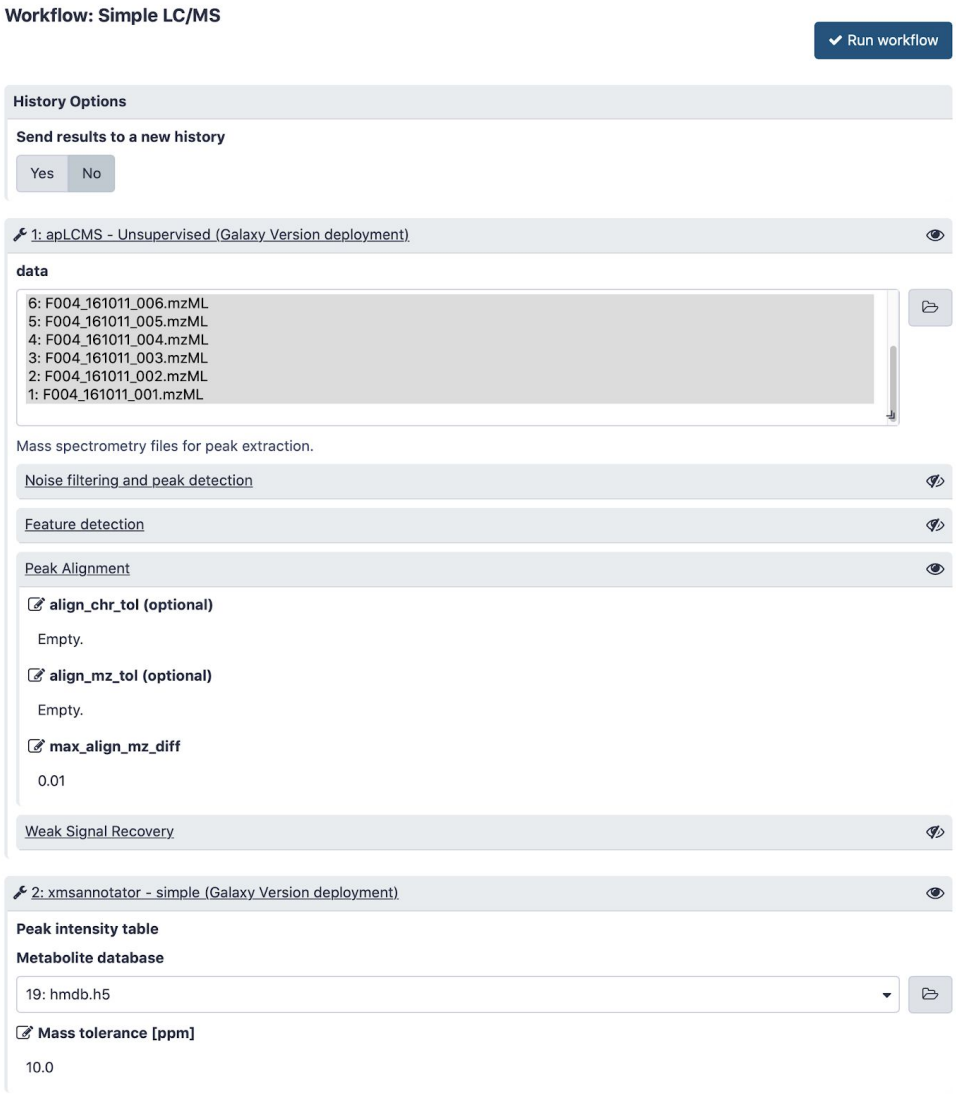


Figure 23. Screenshot of an input form of a simple workflow in UMSA.

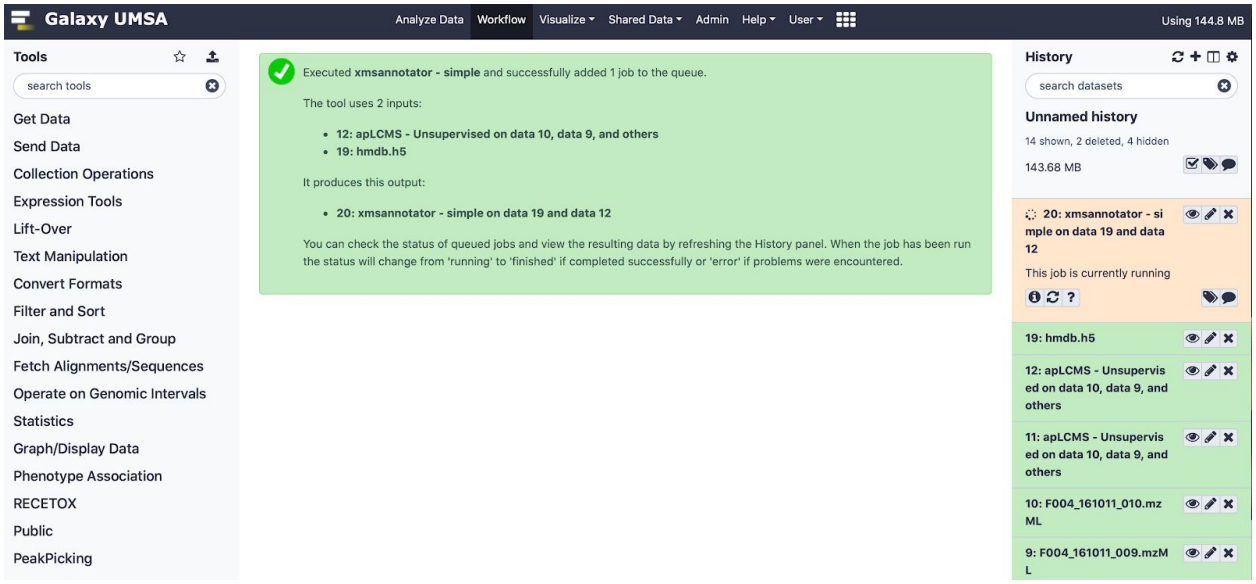


Figure 24. Screenshot of the execution of the previous workflow in progress. The peak-picking step (apLCMS) has already finished, annotation is running.

3.8.5. Demonstration Video

The demonstration video preparation was delayed due to the second outbreak of COVID-19 in the fall of 2020. It is expected to appear in early 2021.

3.9. Modeling and Analysis of Water Supply Systems (MSWSS)

3.9.1. Description

MSWSS is a service for modeling and analysis of Water Supply Systems which integrates the analysis of toxics in drinking-water supply networks with water distribution network simulation. MSWSS service will allow water infrastructure operators and researchers to analyse hazardous events (e.g. toxics propagation within a pipe system) and may be used for preparation of risk management plans for water utilities. The EOSC computing infrastructure and data sharing services enable modelling more complex water supply systems and to increase the number of scenarios for the analysis.

3.9.2. Architecture

The architecture of the MSWSS service is depicted in figure 25. The MSWSS service uses Galaxy portal where users can share and reuse their workflows with data and prepare their simulation jobs.

The computational backend of the MSWSS service is based on an elastic virtual cluster built by EC3 service and managed by the CLUES service and the Infrastructure Manager. EGI Check-in integration allows the use of computational resources available in the EOSC Cloud IaaS infrastructure. The resources inside the virtual cluster are managed by the Slurm workload management system. The data produced by computational jobs are stored locally within the MSWSS service and are available to users via Galaxy portal.

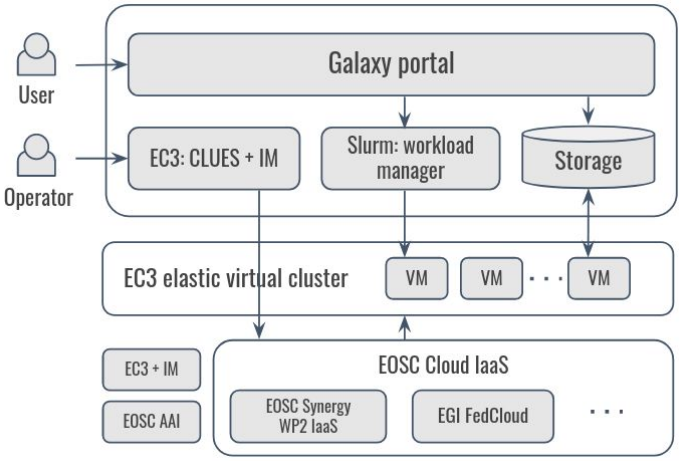


Figure 25. Architecture of the MSWSS Thematic Service.

3.9.3. EOSC Services

MSWSS thematic service has selected and it is integrating the following services listed in the EOSC marketplace:

- EC3 (Infrastructure Manager, CLUES): is used for creation and management of computational backend based on elastic virtual cluster built from virtual worker nodes
- EOSC Cloud computing resources: are used to build the elastic virtual cluster for MSWSS service

- EGI Check-in: it is used by EC3 to authenticate the MSWSS service to EOSC Cloud computing resources

3.9.4. Service Endpoint

The first prototype of the MSWSS service is available at the link https://mswss.ui.savba.sk:8443. Figure 26 shows the main interface of the MSWSS service which is based on a customised Galaxy portal.

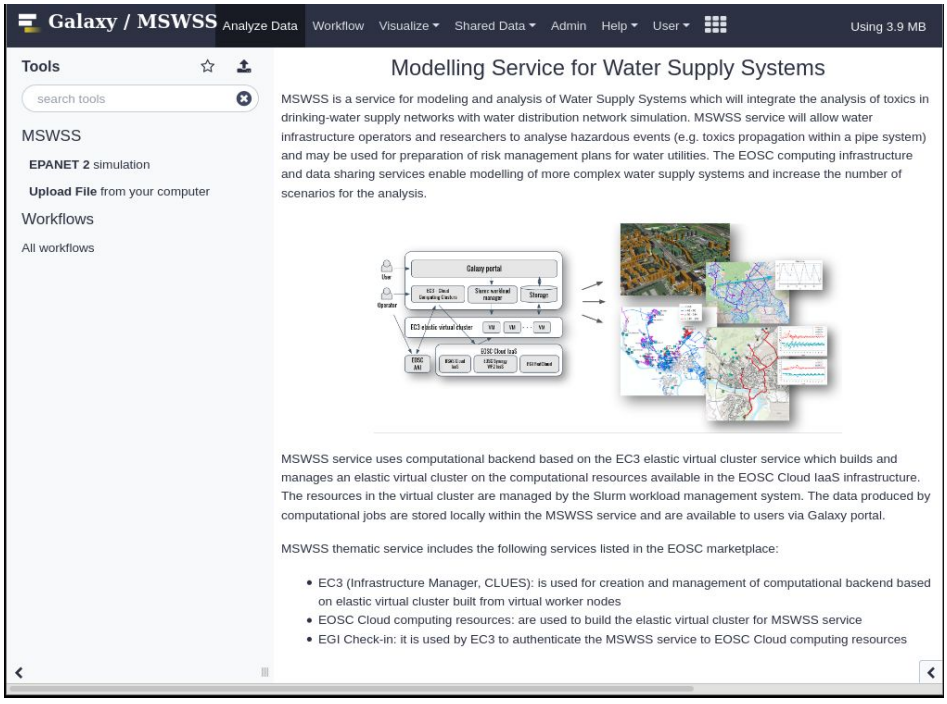


Figure 26. Screenshot of the main interface of the MSWSS Thematic Service

After successful login to the service users can upload the input files and submit the simulation to the workload manager of the elastic cluster which runs their jobs in the EOSC Cloud compute resources. Figure 27 shows the EPANET 2 simulation submitted to the workload manager.

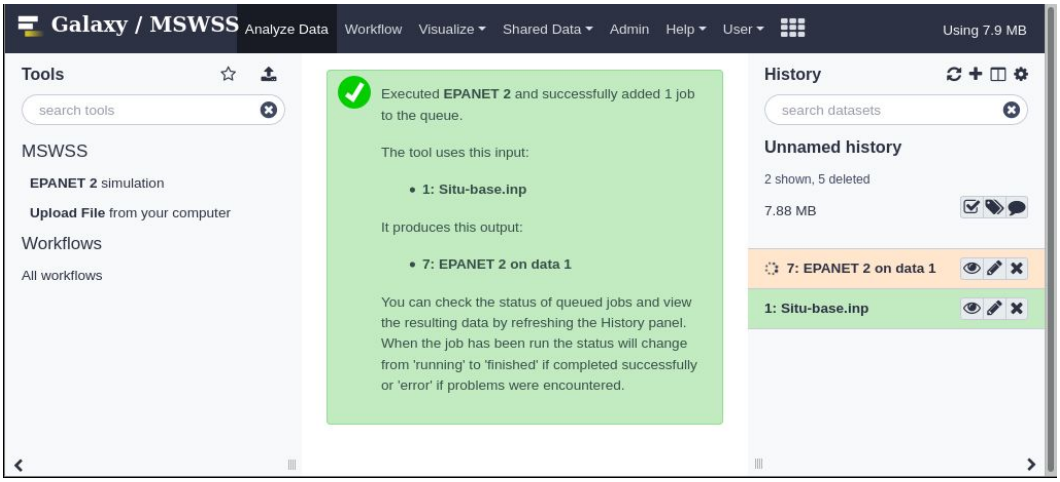


Figure 27. Screenshot of the EPANET 2 simulation job submitted to the EC3 cluster.

The first prototype of MSWSS service provides only basic functionality, more features are planned to be added after MS18, including the ability to prepare and process parametric simulations consisting of large number (tens of thousands) of jobs, post-processing of the results and the integration with EOSC data repositories (OneData).

3.9.5. Demonstration Video

The video presents the first prototype of the MSWSS service. It consists of two parts. First part presents the objective and motivation of the thematic service with the list of the EOSC services integrated. The second part demonstrates the interaction with the service from the point of view of service operators and users. The video is available at the following link: https://youtu.be/ZpKzvUJqZHI

3.10. O3AS

3.10.1. Description

The assessment of ozone depletion is an important task for Climate and Environment studies because the ozone layer in the stratosphere plays an important role for the protection from harmful UV radiation. After the discovery of the ozone hole in 1979, scientists all over the world got worried about the resulting dangers. Soon after, substances that destroy the ozone layer were prohibited with the adoption of the Montreal Protocol, which got ratified unanimously by the United Nations in 1986. Ever since then, reports have been compiled to assess ozone depletion. The last assessment report was published in 2018, the next one is due in 2022. For the compilation of these assessment reports large amounts of climate model projections are analysed to produce time series of ozone to detect trends and possible dates when the ozone gets back to the same level as in 1980. To establish a reliable analysis of the ozone return dates climate scientists need a tool that will help to generate and reliably reproduce figures of high quality. The Ozone assessment service (O3as) is developed within the European Open Science Cloud (EOSC)-Synergy project to assist scientists to visualise ozone data from large climate models. The climate model output is mainly from the Chemistry-Climate Model Initiative (CCMI) project (http://blogs.reading.ac.uk/ccmi/ccmi-phase-two/), with daily model output from 1960 to 2100 of about 20 different models. The recent 2018 quadrennial global assessment of ozone depletion (https://www.esrl.noaa.gov/csl/assessments/ozone/2018/) consists of six chapters and five appendices with about 25 people actively working on each chapter and a multitude of people working in support of the preparation of the document. The service aims to analyze available ozone data from climate models and reanalysis data, calculate return dates for the recovery of the ozone layer and trends of the amount of ozone in the atmosphere to produce results in the form of figures in publication quality. Large volumes of data (TBs) are processed in the complex workflow to generate key metrics.

3.10.2. Architecture

Figure 28 shows the architecture of the O3AS technical service. O3as service is split in several actions and components:

1. A user configures his/her request in the Web Application (currently in development).

2. This request is passed to O3as service via the O3as REST API call.

3. O3as service processes the request, where the pre-processed data (aka skimmed data) is accessed via WebDav and OIDC.

4. In order to produce skimmed data, regular tasks run on HPC to copy primary data and perform data preparation (e.g. data reduction and parameter unification).

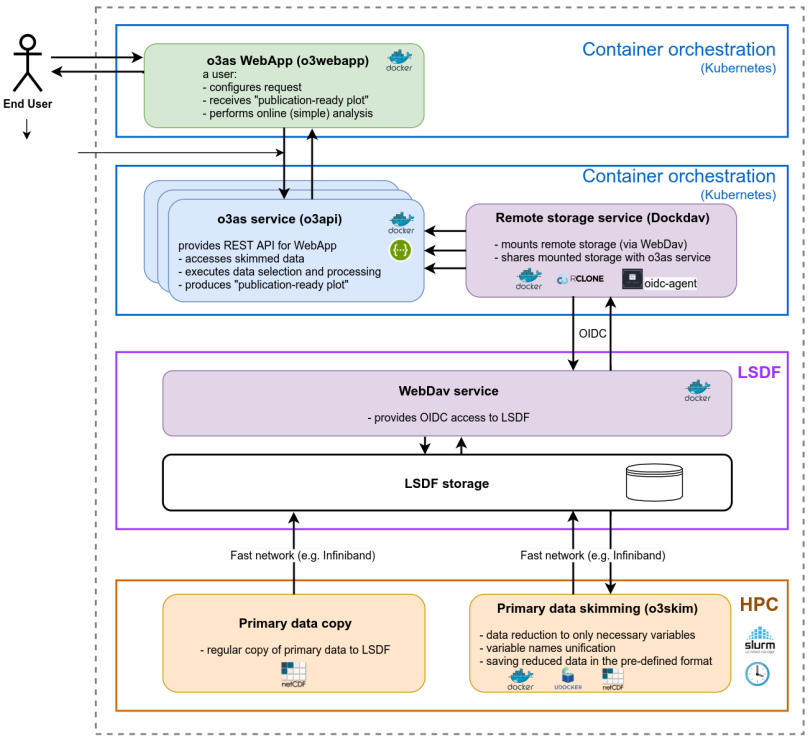


Figure 28. Architecture of the O3AS Thematic Service.

3.10.3. EOSC Services

The O3as thematic service depends on several services that are provided by EOSC.The EOSC OIDC EGI Check-in shall be used to access the advanced features of the service, i.e. the plotting of advanced figures and saving the plotted data points for further analysis. Under the hood to access data on the data server LSDF the OIDC-agent is used for mounting using WebDAV-HTTP authentication. In order to perform the data reduction/skimming in the HPC environment the EOSC development udocker is a vital part for running containers in environments where root access is not available. Finally for the deployment of the service resources, the Infrastructure Manager is used to start Kubernetes cluster at one of the EOSC cloud providers for the project's Docker containers . To summarize O3as selected and it is integrating the following services listed in the EOSC marketplace:

- EOSC OIDC providers, e.g. EGI Check-In: to access certain functionalities of the service.
- Infrastructure Manager: To deploy service components on EOSC cloud resources.
- EOSC cloud resources: to offer O3as service for users in a timely manner.

A part from these services we also use the following tools developed in EOSC-related projects:

- OIDC-Agent: To mount data servers using WebDAV - HTTP authentication.
- udocker: To perform data skimming in the HPC environment.

3.10.4. Service Endpoint

The O3as service entry point is o3as.data.kit.edu. The portal provides basic information about the service, contacts, where to look for further detailed documentation, source code, Docker Hub, and, of course, link to the deployed service (figure 29).

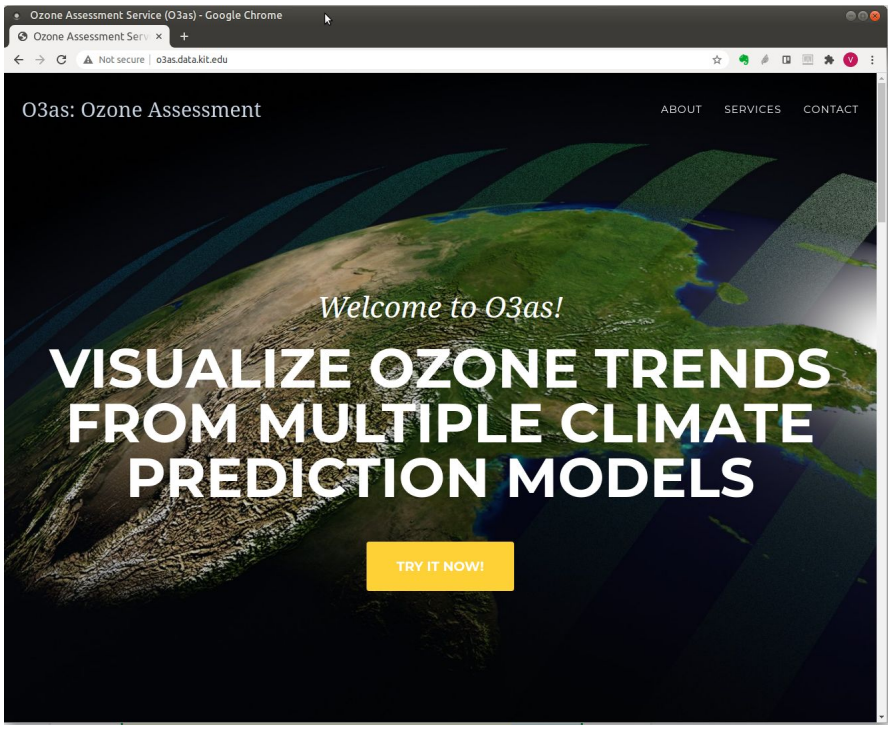


figure 29: O3as entry portal.

It currently directs a user to the o3api component (figure 30), which accepts user requests and delivers plots of interest for Ozone assessment. It runs in the Kubernetes cluster deployed at one of the EOSC-SYNERGY cloud providers (CESGA) via the Infrastructure Manager. One other component, o3skim runs regularly on the HPC system of KIT, ForHLR2, to reduce the raw Ozone data to the parameters of interest for the Ozone assessment. The reduced data are made available to o3api. In the ongoing development, beyond MS18, we will add another part of the service, the web application (o3webapp), where users may customise their requests in a more friendly way. Researchers, who are members of o3as.data.kit.edu Virtual Organization, after authenticating via EGI Check-In will get advanced functionality of the service (e.g. receiving data points corresponding to the request, perform basic analysis of the data).

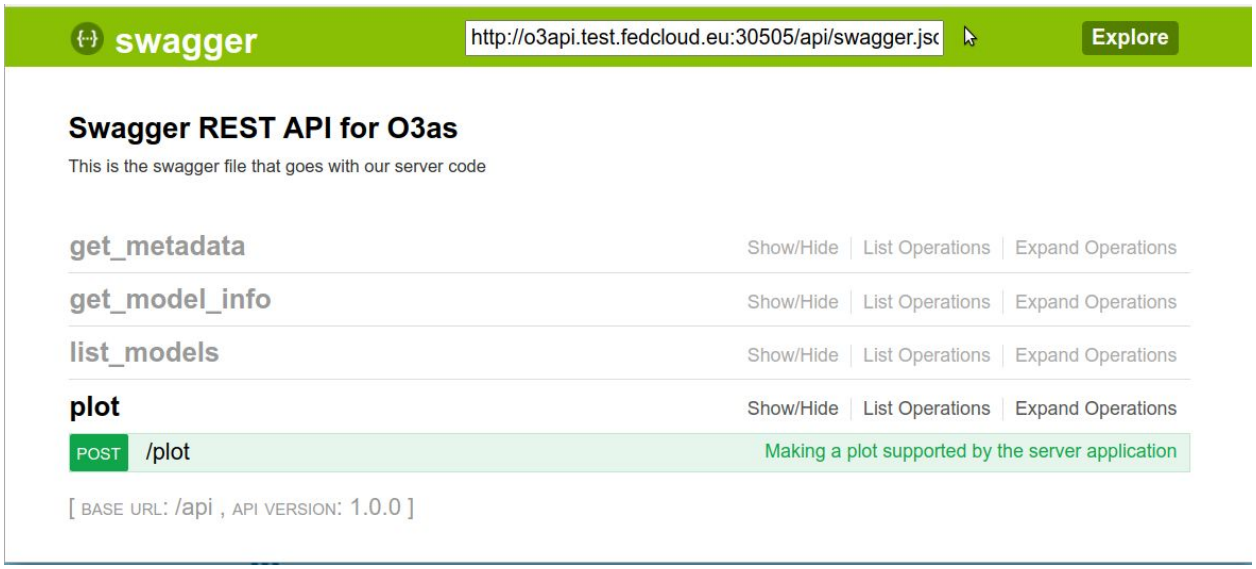


figure 30: o3api component is ready to accept user requests and deliver scientific plots for Ozone assessment.

3.10.5. Demonstration Video

The video presentation of O3as service, which can be found at the link below, describes the problem climate scientists deal with in the case of Ozone assessment, and the solution offered by O3as service. It demonstrates the current status of the service, how one can access it and use it, and planned upgrades. In particular, we show how the original Ozone raw data are reduced to the parameters of interest, how the service can be deployed by the means of the Infrastructure manager and started in the Kubernetes cluster, and what functionality is available for users, e.g. how to configure the requests and receive plots of interest. The video also presents public weblinks for further exploration of the service.

Link to the video: https://youtu.be/mQo7y3tyX08

4. Conclusion

This report supports the demonstration deliverable D4.2 - First prototype of the EOSC Thematic services which shows four operational EOSC-SYNERGY thematic services although with limited functionality. The deliverable shows additional information about the status and plans of all the thematic services.

All thematic services have identified several EOSC technical services to address some of the challenges and requirements that were not properly fulfilled in the initial versions. Each thematic service has differences that led to the adoption of one or another thematic service, which enriches the catalogue of experiences, best practices and solutions. As a summary, three different (although compatible among them) AAI methods have been integrated (EGI Checkin, B2ACCESS and Life-Sciences AAI). Job scheduling ranges from solutions based on containers (using Kubernetes) to solutions using batch queues (mainly based on SLURM), supported in some cases by workflow frameworks such as Galaxy and instantiated through EC3. For the interaction with cloud resources, TOSCA and RADL recipes have been developed for Infrastructure Manager. Data access is performed through different solutions such as Dataverse, WebDav, EGI Datahub OneData, B2SHARE and B2SAFE, which clearly states the complexity of the data management issue and the wide range of solutions.

The experience among the thematic services will be extremely useful for new services to be developed, so an important effort on communication will be performed. Opportunities such as the EOSC-hub week will be used to showcase demonstrations and presentations.