Thematic Service integration in EGI Fedcloud

Amanda Calatrava
On behalf of WP4

EGI Conference 2022

Agenda

0. Introduction: the 10 thematic services
1. Increasing the capacity, performance, reliability and functionality
2. Increasing service quality
3. Increasing relevance of National Thematic Services
4. Increasing the number of users
5. Conclusions

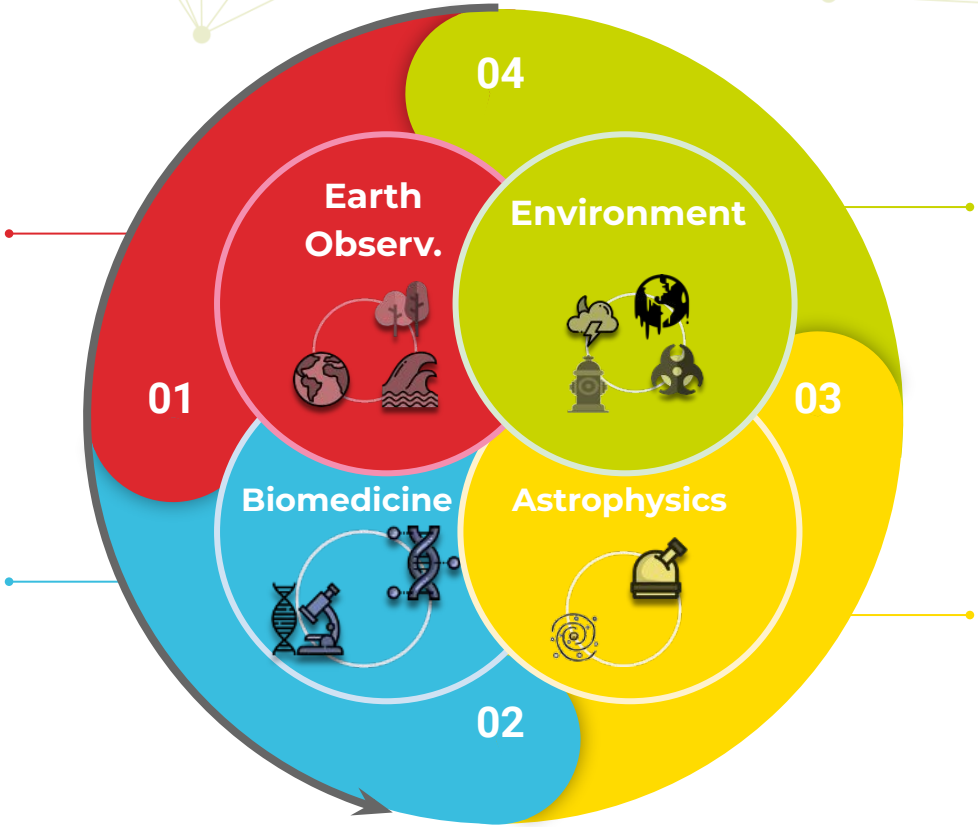
0. The Thematic Services Virtuous Cycle

Increase the capacity, performance, reliability and/or functionality

By means of best practices for adopting common EOSC core tools and services.

Increase service quality

FAIR data practices and software quality assessment.



Increase relevance of National Thematic Services

By expanding the use of the mature national services in an international scope.

Increase the number of users

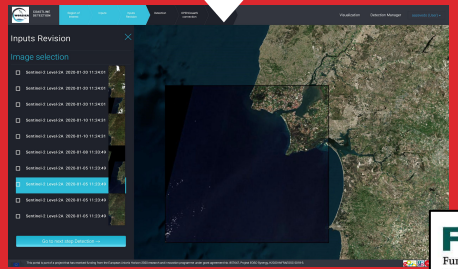
By means of the integration in EOSC and the training.

0. Thematic Services in Earth Observation

WORSICA

Water Monitoring Sentinel Cloud Platform

A service for the detection of water using satellites, Unmanned Aerial Vehicles & in-situ data. WORSICA can be used for coastline detection, inland water bodies detection and water leaks detection on irrigation networks.



SAPS

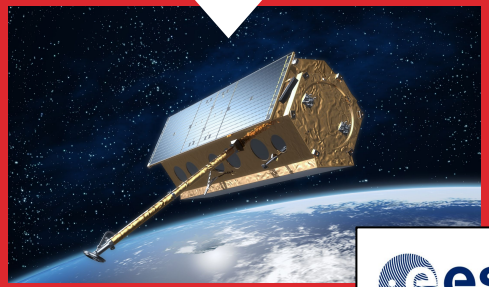
Surface Energy Balance Automated Processing Service

Used to estimate Evapotranspiration and other environmental data that can be applied, for example, on water management and the analysis of the evolution of forest masses and crops.

GCore

Acquisition, cataloguing and processing EOS data

G-Core is a production-ready technology used as a service at ESA's and national programs that provides a Data Manager for spatial and non-spatial purposes and a framework for third-party processors.



0. Thematic Services in Biomedicine & Astrophysics

SCIPION

CryoEM data processing for Structural Biology

ScipionCloud service will allow users from Instruct to deploy a dynamic cluster in the cloud to keep processing the data acquired at the facility.

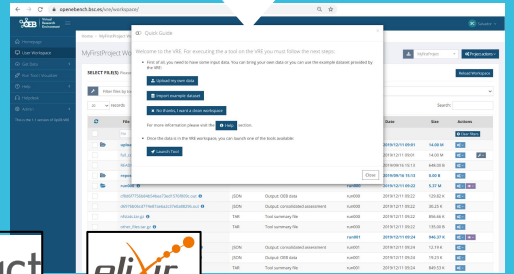
EIRENE

instruct ERIC

OpenEBench

ELIXIR benchmarking and technical monitoring platform

Used to evaluate bioinformatics tools, OpenEBench is an observatory for SW quality based on the automated monitoring of FAIR for research software metrics and indicators.



elixir

LAGO

Latin American Giant cosmic ray Observatory

LAGO is a cosmic ray observatory made of a network of water- Cherenkov detectors (WCD) spanning over different altitudes and latitudes making research on High Energy Physics, Space weather, etc.

LAGO

0. Thematic Services in Environment

UMSA

Untargeted Mass-Spectrometry Analysis

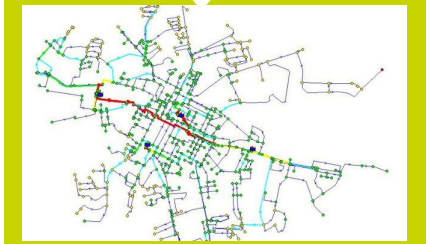
UMSA aims at processing data to correlating the whole spectra with other data to work with more complex hypotheses on the impact of environment in human health.

EIRENE

MSWSS

Water Supply Systems modeling and analysis

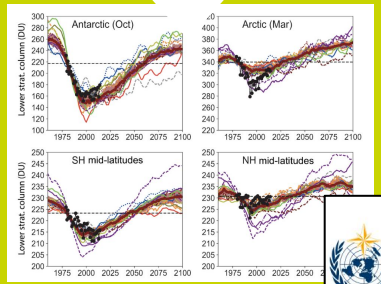
MSWSS integrates the analysis and simulation of toxics in drinking water supply networks to allow operators and researchers to analyse hazardous events.



O3AS

Ozone Analysis Service

The O3AS service shall provide an invaluable tool to extract O_3 trends from large climate prediction model data to produce figures of stratospheric ozone trends.

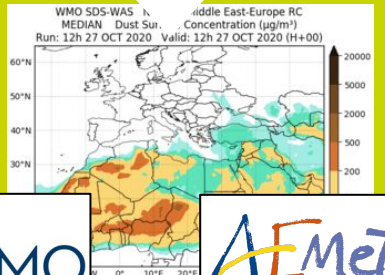


WMO

SDS-WAS

A Service related to the mineral dust forecast

SDS-WAS aims to support institutional entities to warn about possible dust events and to foster the study of dust-related phenomena.



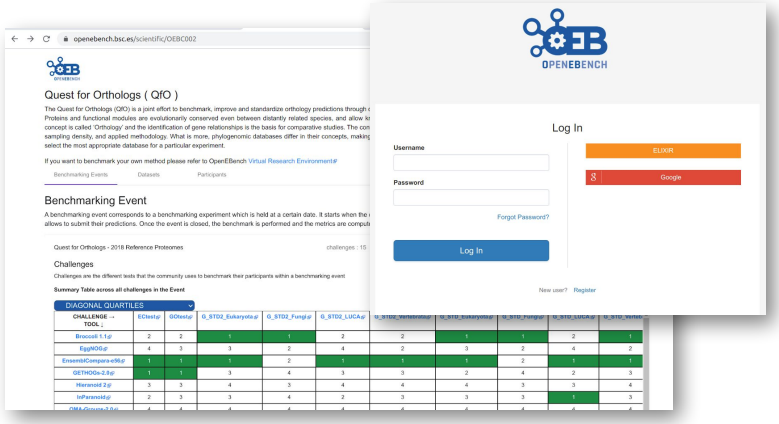
AEMet
Agencia Estatal de Meteorología

1. Increasing the capacity, performance, reliability and functionality: Analysis of Gaps and Bottlenecks

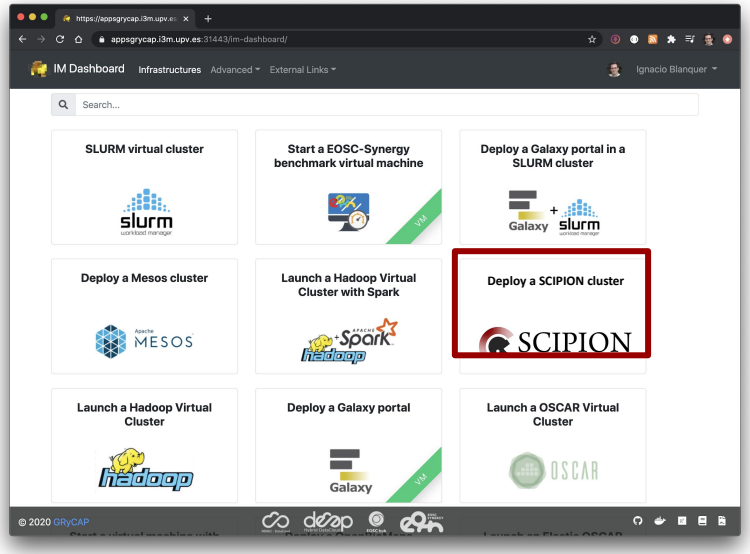
DM	**Data Management**	• Ensure bandwidth for downloading data from operational providers • Need for Persistent Data Storage for resulting data with POSIX interface. • Data access control mechanisms. • Need of fast catalogues for data discovery.
WM	**Workload Management**	• Need for processing resources. • Need for containerised workloads.
DP	**Processing Resources**	• Need for Resource Management Services. • Non-standard configuration (RAM and GPU). • Dynamic infrastructures.
AAI	**Authentication and Authz Infrastructure**	• Use of federated Identity Providers. • Identity Delegation.

1. Increasing the capacity, performance, reliability and functionality: different usage models

- A single entry portal that manages users, data access, processing and visualization
  - Supported by a shared or dedicated pool of resources.



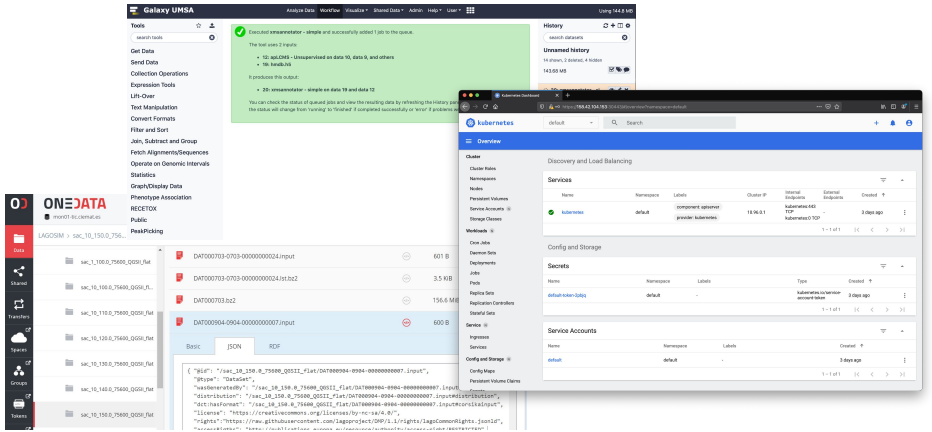
- Each user deploys his/her own instance
  - A combination of TOSCA recipes with Docker containers.



1. Increasing the capacity, performance, reliability and functionality: different Computing Backends, Workloads & Storage

- A dynamic dedicated cloud backend
  - An elastic cluster that shrinks or grows according to the workload.
- Supported by Kubernetes, SLURM Batch queues or Galaxy.
- Links to external HPC (Marenostrum) and HTC (EGI HTC compute) for massive Batch jobs.

- Integrating external Data Infrastructures
  - Both EGI DataHub and EUDAT.
- Deploying their own Datastore
  - E.g. DATAVERSE instance.



1. Increasing the capacity, performance, reliability and functionality: Adoption of EOSC Services

Service	WORSICA	G-Core	SAPS	Scipion	OpenEBench	LAGO	SDS-WAS	UMSA	MSWSS	O3AS
AAI	***EGI Check in***	CAS User/pwd & ***EGI Check in***	***EGI Check in***	***EGI Check in***	Life Sciences AAI	***eduTEAMS+ EGI Check-in***	***B2ACCESS***	***EGI Check in*** & Life- science AAI	***EGI Check in***	***EGI Check in***
Workload Mng.	ArcCE, Batch (SLURM)	GCore+ K8s	K8s	Batch (SLURM)	WfExS + NextFlow	Batch (SLURM)	Batch (SLURM)	Batch (SLURM) in ***IM/EC3*** (in Galaxy)	Batch (SLURM) + Galaxy	Cluster batch (SLURM) & K8s
Resource Mng.	***IM (TOSCA)***	***IM***	***EC3***	***IM / EC3***	Opennebula	***Local clusters + IM***	Local clusters	***IM / EC3***	***EC3***	Local cluster + **IM**
Data Storage	Nextcloud, Dataverse	ElasticSearch	OpenStack Swift	Local + ***EGI Datahub***	Local + ***B2SHARE***	***EGI DataHub ONEDATA***	***B2HANDLE /B2SAFE***	Local + S3	Local + Dataverse	Local + WebDAV

D4.3.-Final release of the EOSC Thematic services https://doi.org/10.20350/digitalCSIC/14611

1. Increasing the capacity, performance, reliability and functionality: Improvements due to EOSC-Synergy

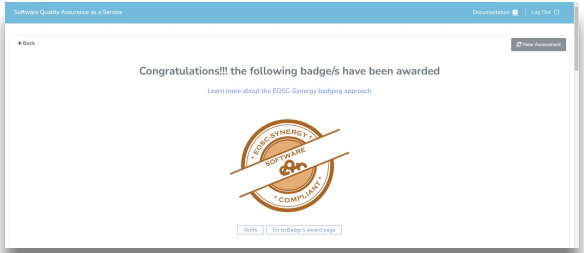
- Integration of **standardized AAI IdPs** to facilitate user management.
- Improvement of **processing backends** by replacing single computing instances with batch job queues, container management platforms or clients to high-throughput computing backends.
- **Publishing** the output results in persistent repositories.
- **PID annotation** of output data and integration in official harvesters.
- Improving **repeatability** and **platform-agnosticism** by describing the application topologies as code using standard TOSCA language.
- **Self-management** of resources to reduce maintenance costs.

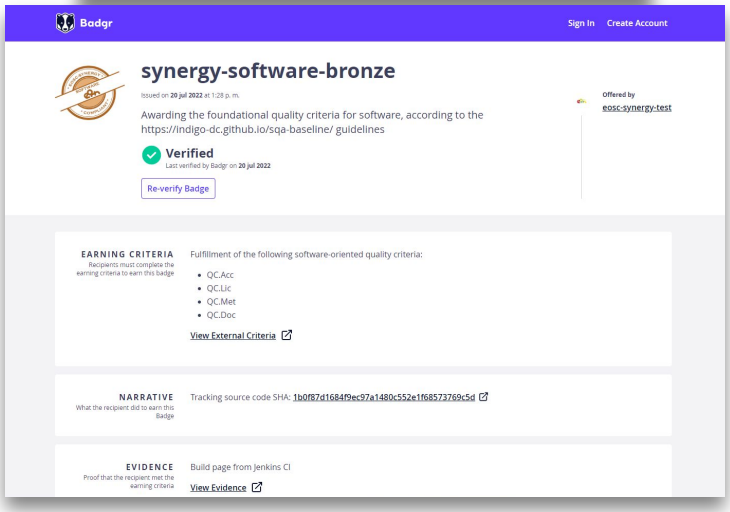
2. Increasing service quality: Software and service quality evaluation

Test Software	• Software Quality Assurance baseline • **Tox** automation tool integrated with **SQAaaS/JePL** and a **Jenkins** server
Deliver Service	• **Build docker images** to automate the deployment • **Deploy** virtual infrastructures automatically
Comply with FAIR Data principles	• **FAIR** principles **recommendations**. • **Framework** to support FAIR best practices: implementation, validation, monitoring

2. Increasing service quality: Software quality evaluation

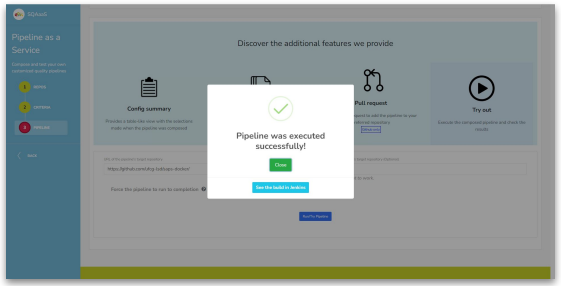
- Ensure **software quality** of the service by following the **DevOps** culture, enabling a **continuous integration** (CI) and **delivery** (CD) approach.
  - The SQAaaS service facilitates the creation of the pipelines.
  - Quality Assessment & Awarding is also recently available, where badges are issued.





2. Increasing service quality: Service quality evaluation

- Facilitate the **delivery** of the services to easily reproduce the architecture deployment by evaluating the building and deployment of the services:
  - Building containers.
  - Deploying a virtual infrastructure with IM or EC3.



	Declarative: Checkout SCM	SQA baseline criterion: QC.Doc & QC.Sty & QC.Uni & QC.Sec	Environment Setup	qc_doc saps-common	qc_security saps-common	qc_style saps-common	qc_coverage saps-common	Push Images to Docker Registry	Docker Compose cleanup
Average stage times: (Average full run time: ~7min 33s)	3s	14s	3s	7s	5min 54s	7s	12s	11s	18s
#28 Sep 10 12:08 No Changes	3s	14s	5s	15s	5min 32s	6s	12s	10s	17s

	Declarative: Checkout SCM	SQA baseline criterion: SvcQC.Dep	Environment Setup	SvcQC.Dep this_repo	Docker Compose cleanup
Average stage times: (Average full run time: ~11min 10s)	2s	13s	2s	10min 16s	16s
#1 Sep 09 12:01 No Changes	2s	13s	2s	10min 16s	16s

2. Increasing service quality: FAIR Data evaluation

- Comply with **FAIR data principles**, allowing the **reusability** of the output data and the **findability** by other researchers.
  - Support in the selection of the data format, assessment with best practices.
  - Define the Data Management Plan (DMP).
  - Build the Jenkins pipeline to do the FAIRness check. Two automated tools:
    - F-UJI (FAIRsFAIR).
    - FAIR Evaluator (DIGITAL.CSIC).

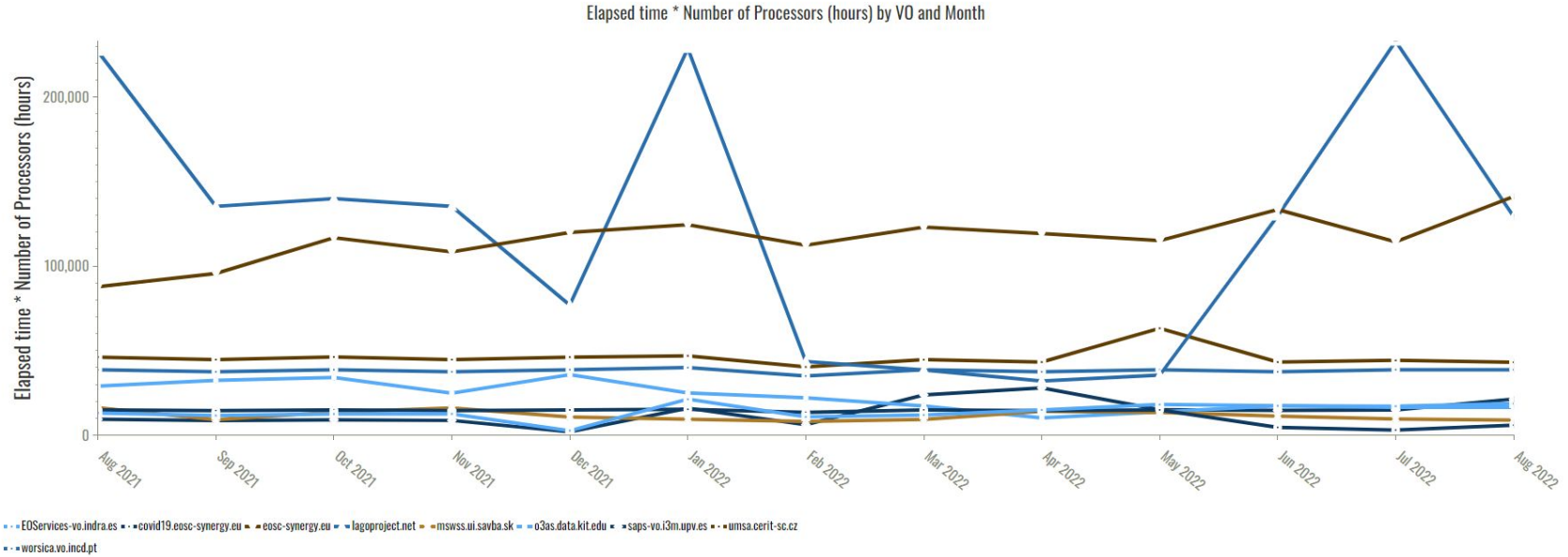
2. Increasing service quality: Thematic Services adoption

	Software Quality Assurance (Tool/Repos)	Service Quality Assurance	FAIR Data principles (Repository)
WORSICA	JePL / 4 (Tox)	Containers on existing K8s	On local Dataverse
SAPS	JePL / 7 (Maven)	Virtual K8s cluster and containers	Processing service over external FAIR Data
gCORE	Private JePL	Virtual K8s cluster and containers	ESA standards
SCIPION	N/A	Virtual SLURM cluster recipe and containers	Processing service over external FAIR Data
OpenEBench	JePL / 1 (Tox)	Existing deployment	B2HANDLE + B2SHARE
LAGO	JePL / 2 (Tox)	virtual SLURM cluster recipe	ONEDATA / B2Handle + B2Find
SDS-WAS	SQaaS	Existing deployment	B2HANDLE + B2SAFE
UMSA	CI/CD GitHub pipelines	Virtual Galaxy+Slurm cluster	On local Repository
MSWSS	N/A	Virtual Galaxy+Slurm elastic cluster	On local Dataverse
O3AS	JePL / 4 (tox)	Containers on existing K8s	On local Repository
	Completed	In Progress	In Progress

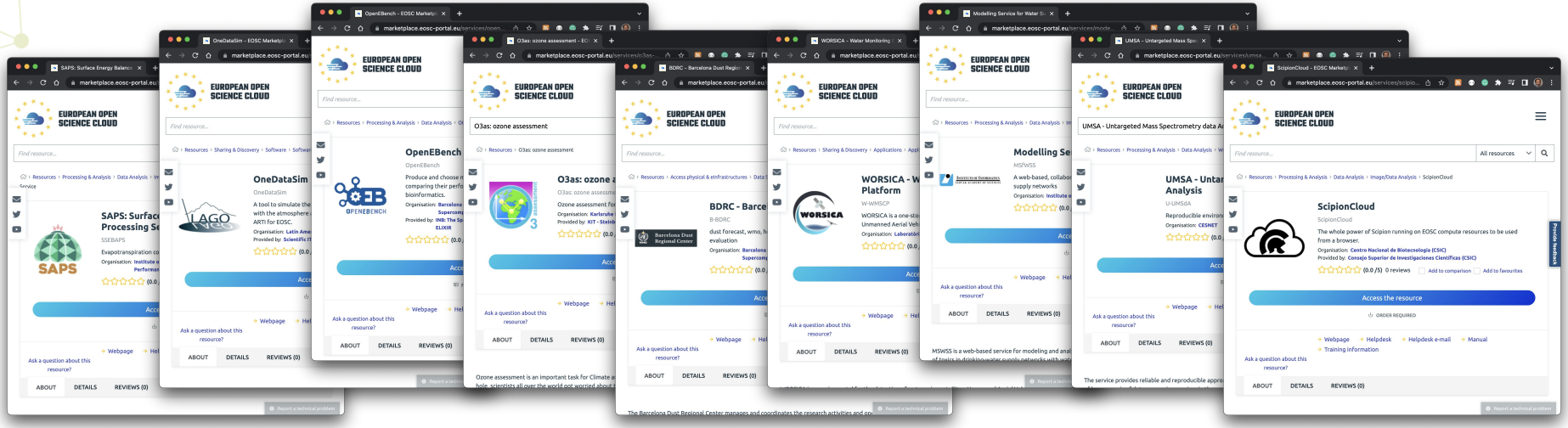
3. Increasing relevance: Measuring success - Metrics and KPIs: Linkage to e-Infrastructures

In the last year:

- Over 4.7 Million CPU core hours
- More than 3.400 VMs
- More than 90 registered users in VOs



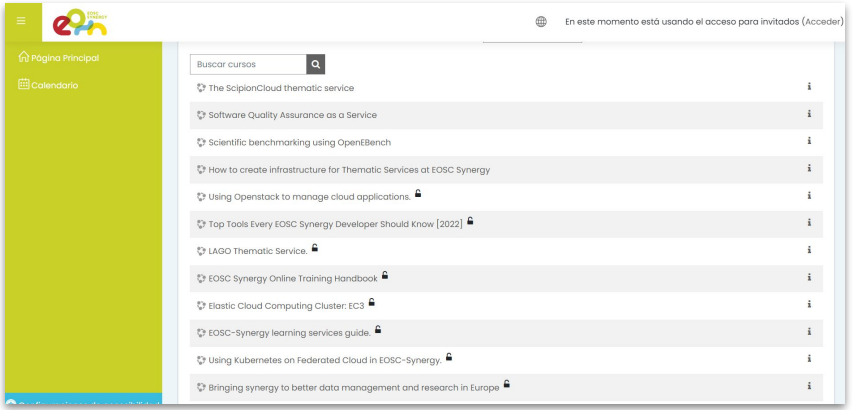
4. Increase the number of users: Integration in EOSC Marketplace

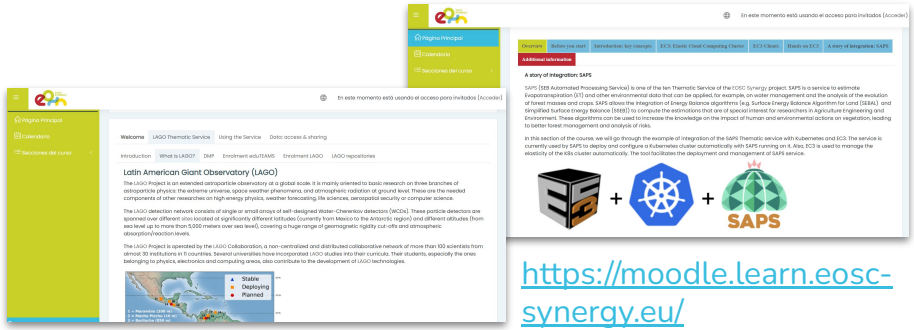


EUROPEAN OPEN SCIENCE CLOUD

4. Increase the number of users: Training materials produced

- Several courses and training materials produced by the TSs:
  - 4 full courses on the moodle platform.
  - contributed to the learning catalog with a wide variety of training materials, including tutorials, documentation and presentations from physical training programs.

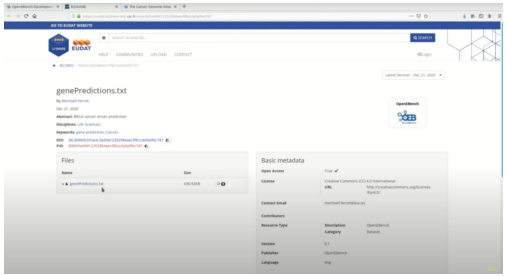


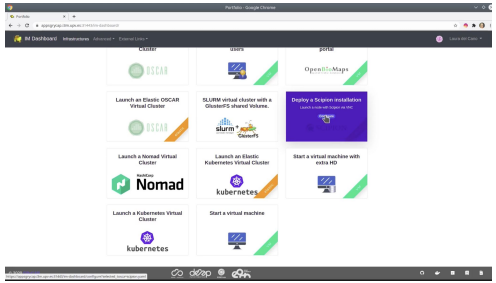


https://moodle.learn.eosc-synergy.eu/

4. Increase the number of users: Demonstration videos

- Examples of the best practices from the videos (https://bit.ly/3bXJZq2)
  - Integration of standardized AAI IdPs (UMSA 1:08)
  - Improvement of processing backends (SAPS 0:50).
  - Publishing the output results in persistent repositories. (SDS-WAS 4:04)
  - PID annotation of output data & integration in official harvesters (OpenEBench 7:20)
  - Improving repeatability and platform-agnosticism through standard TOSCA descriptions (SCIPION 3:15).
  - Self-management of resources to reduce maintenance costs (MSWSS 1:55).





5. Conclusions

- EOSC-SYNERGY aims at Building Capacities in EOSC through the development of ten data-intensive thematic services oriented to different scientific disciplines.
- The adaptation, improvement and quality assessment of those services on a Federated Data Infrastructure strongly aligns with the objectives of EOSC (*) and will develop best practices and experiences
  - A key factor for the success of EOSC (**) is performance: how EOSC as an ecosystem operates and how the resources are used and acknowledged by the users.
  - All the services consume services from the EOSC catalogue, which will provide feedback on the usability and relevance of the model.

(*) Draft EOSC partnership proposal: "..It aims to accelerate the deployment and consolidation of an open, trusted, virtual, federated environment in Europe to store, share and re-use research data across borders and scientific disciplines and provide access to rich array of related services..".

(**) Solutions for a Sustainable EOSC: An Iron Lady report from the EOSC Sustainability Working Group, Draft 16 September 2020.

Questions & Contact

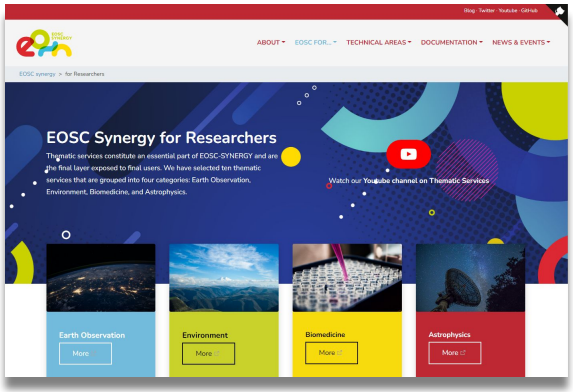
Amanda Calatrava Arroyo

amcaar@i3m.upv.es

Instituto de Instrumentación Para la Imagen Molecular

Universitat Politècnica de València

Gracias !
Obrigado !
Danke !
Dziękuję !
Vďaka !
Děkuji !
Bedankt !
Merci !
Thanks !



www.eosc-synergy.eu

@EOSC_synergy