D7.5 Final report on Thematic service architecture, software integration and exploitation

Lead Partner:	CINECA
Version:	1
Status:	Under EC review
Dissemination Level:	Public
Document Link:	https://documents.egi.eu/document/3641

Deliverable Abstract
This document provides the description of the final releases and achieved integration for the thematic services part of EOSC-hub WP7 activities. For each thematic service, it also describes the impact, lesson learnt, exploitation of the services and how it will be sustained after the end of EOSC-hub.

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

COPYRIGHT NOTICE

DELIVERY SLIP

Date	*Name*	*Partner/Activity*	*Date*
From:	Debora Testi	CINECA/WP7	01/04/2021
Moderated by:	Malgorzata Krakowian	EGI Foundation/WP1	
Reviewed by:	Gergely Sipos Matti Heikkurinen	EGI Foundation/WP8 EGI Foundation/WP3	31/03/2021
Approved by:	AMB		

DOCUMENT LOG

Issue	*Date*	*Comment*	*Author*
v.0.1	16/12/2020	Table of content	Debora Testi (WP7 Leader)
v.0.2	19/02/2021	Full draft for WP review	Willem Elbers (CLARIN), Dieter Van Uytvanck (CLARIN), Daniele Spiga (INFN), Fabrizio Antonio (CMCC), Alessandro D'Anca (CMCC), Donatello Elia (CMCC), Stephan Kindermann (DKRZ), Mattia Santoro (CNR), Anabela Oliveira (LNEC), Alexandre Bonvin (UU), Antonio Rosato (CIRMMP), Philipp Wieder (GWDG), Davor Davidovic (IRB), Sheila Izquieta Rojano, Juan Miguel González-Aranda (LW-ERIC), Antonio José Sáenz-Albanés (LW-ERIC)
v.0.3	31/03/2021	External review completed	Gergely Sipos, Matti Heikkurinen
v.1	01/04/2021	Final	Debora Testi (WP7 Leader)

TERMINOLOGY

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

Contents

Executive summary

The research communities, which are partners in WP7, are both service consumers and providers. They offer services to their users, which are called Thematic Services (TSs). Thematic Services are scientific services (incl. data) that provide discipline-specific capabilities for researchers (e.g. browsing and downloading data and apps, workflow development, execution, online analytics, result visualisation, sharing of result data, publications, applications). In some cases, in order to integrate those services with EOSC-hub (which is the main scope of WP7), the software code implementing them has been modified, by extending existing components or developing new plugins.

WP7 involves the participation of 9 Thematic services from different scientific domains with the aim to achieve the integration of core services in communities services or workflows and to make available thematic services to the scientific communities:

- CLARIN: component metadata framework (CMDI) as a way to stimulate the discoverability of data sets, especially in the humanities and social sciences.
- DODAS: Dynamic On Demand Analysis Service.
- ECAS: ENES Climate Analytics Service.
- GEOSS: Global Earth Observation System of Systems
- OPENCoastS: On-demand oPEratioNal Coastal circulation forecast Services.
- WeNMR: Worldwide e-Infrastructure for NMR and structural biology.
- EO Pillar: a set of services related to Earth Observation (EO) domain.
- DARIAH: Digital Research Infrastructure for the Arts and Humanities
- LifeWatch: a set of services related to Biodiversity and Ecosystem.

Due to the specificity of the requirements, each Thematic service community has worked independently from the others, but experiences and issues were shared among the different teams during the lifetime of the project. Thus, in addition to performing the integration work more efficiently, EOSC-hub has created conditions that increase the potential for cross-pollination between the communities as a whole (e.g. through reuse of tools and resources across disciplines, supported by technical staff with skills extending beyond the solutions of a specific discipline) in the exploitation phase of the project.

The software architectures have been described in detail in a previous document (D7.4), while this deliverable will provide any recent technical update and information on the impact of the achieved integrations together with some lessons learnt from the Thematic services.

At the end of EOSC-hub, all TSs have one or more services exposed in the EOSC marketplace and showed an increased usage and engagement with the respective communities.

1 Introduction

This deliverable provides the overview of the final results of the integration activities of the Thematics services part of EOSC-hub WP7.

The document is organised with a section for each Thematic service including a similar set of information for all services:

- General overview of the services
- Ambition at the start of the project in 2018
- Final software architecture and integration
- Impact and exploitation
- Lesson learnt and future plans beyond EOSC-hub

EO Pillar Thematic service is not included in this document as the related activities have been completed in the first year of the project and the integration description has been provided in previous deliverables of WP7.

2 CLARIN

2.1 Service description

Service/Tool name	**Virtual Language Observatory**
Service/Tool url	https://vlo.clarin.eu
Service/Tool information page	https://www.clarin.eu/vlo
Description	A facet[1] browser for fast navigation and searching in huge amounts of metadata.
Value proposition	A facet browser for fast navigation and searching in large amounts of metadata. This portal enables the discovery of language data and tools, provided by over 40 CLARIN centres, other language resource providers and Europeans.
Customer of the service/tool	Researchers Repository managers
User of the service/tool	Researchers
User Documentation	https://vlo.clarin.eu/help
Technical Documentation	https://trac.clarin.eu/wiki/CmdiVirtualLanguageObservatory (requires authentication, information available on request)
Product team	CLARIN ERIC
License	GPLv3
Source code	https://github.com/clarin-eric/VLO
Testing	Unit and integration tests using junit Vulnerability analysis provided by snyk.io

Service/Tool name	**Virtual Collection Registry**
Service/Tool url	https://collections.clarin.eu
Service/Tool information page	https://www.clarin.eu/content/virtual-collections
Description	A service that allows researchers to create their own citable digital bookmarks.
Value proposition	A virtual collection is a coherent set of links to digital objects (e.g. annotated text, video) that can be easily created, accessed and cited. The links can originate from different archives, hence the term virtual.

[1] Faceted search is a technique which involves augmenting traditional search techniques with a faceted navigation system, allowing users to narrow down search results by applying multiple filters based on classification of the items.

Customer of the service/tool	Researchers
User of the service/tool	Researchers
User Documentation	https://collections.clarin.eu/help
Technical Documentation	https://trac.clarin.eu/wiki/VirtualCollectionRegistry (requires authentication, information available on request)
Product team	CLARIN ERIC
License	GPLv3
Source code	https://github.com/clarin-eric/VirtualCollectionRegistry
Testing	Unit and integration tests using junit Vulnerability analysis provided by snyk.io

Service/Tool name	**Language Resource Switchboard**
Service/Tool url	https://switchboard.clarin.eu
Service/Tool information page	https://www.clarin.eu/content/language-resource-switchboard
Description	A web application that suggests language analysis tools for specific data sets.
Value proposition	The Language Resource Switchboard will automatically provide a list of available tools, based on the language and format of the input. The Switchboard can also be invoked from the Virtual Language Observatory and B2DROP.
Customer of the service/tool	Researchers Tool administrators
User of the service/tool	Researchers
User Documentation	https://switchboard.clarin.eu/help
Technical Documentation	https://github.com/clarin-eric/switchboard-doc/
Product team	CLARIN ERIC
License	GPLv3
Source code	https://github.com/clarin-eric/switchboard
Testing	Unit and integration tests using junit Vulnerability analysis provided by snyk.io

CLARIN is offering three thematic services within the EOSC-hub: The Virtual Language Observatory (VLO), the Virtual Collection Registry (VCR), and the Language Resource Switchboard (LRS).

- A detailed description for the VLO service has already been described in D7.1 and D7.2 [R1], [R2].
- The VCR is a registry provided by CLARIN where scholars can create and publish virtual collections. A virtual collection is a coherent set of links to digital objects (e.g. annotated text, video) that can be easily created, accessed, and cited. The links can originate from different archives, hence the term virtual. A virtual collection is suitable for manual access (using a web-browser) as well as automated processing (e.g. by a web service). The VCR is closely integrated with the CLARIN federated authentication and authorization infrastructure and provides persistent identifiers for easy citation. The collection metadata is openly available and accessible via the VLO.
- The LRS is a tool that helps to find a matching language processing web application for data. After uploading a file, providing a persistent identifier or entering a URL, the resource referenced is analysed and the user is provided a list of tasks he/she can perform on the specified resource. After selecting a task, the user is provided with a list of available CLARIN tools to analyse the input. By selecting a tool, the LRS will make sure the user is sent to the service and the service is properly instrumented with the provided resource.

2.2 Initial ambition (in 2018)

Our initial ambition back in 2018 was described in the EOSC-hub task 7.1 roadmap[2]. For all three thematic services this included iterative releases to improve the individual services, integration into the EOSC-hub marketplace, implementing extensions to link with the EOSC-hub virtual access reporting framework and work on service specific improvements. These were defined as follows: for VLO: improve the integration between the VLO and the Switchboard and the VLO and B2FIND, offer community specific VLO deployments based on requests via the marketplace; for VCR: improving the integration endpoint and adding support for these improvements into the VLO integration with the VCR. Add VCR integration in B2SHARE; for Switchboard: improve the integration with B2DROP and add integration with B2SHARE.

2.3 Final software architecture and integration

The technical software architecture of the different services has already been described in detail in the earlier deliverables, VLO in D7.1, VCR and Switchboard in D7.3. This architecture remained stable until the end of the project, but all thematics services have been continuously improved and updated. Versions released under EOSC-hub, focusing on the planned integrations:

* VLO[3]: 4.6.0, 4.7.0, 4.8.0, 4.9.0 and 4.10.0 before the end of the project.

* VCR[4]: 1.3.0, 1.4.0 and 1.5.0

[2] https://office.clarin.eu/v/CE-2018-1175-EOSC-hub-task71-roadmap.pdf
[3] https://github.com/clarin-eric/VLO/releases
[4] https://github.com/clarin-eric/VirtualCollectionRegistry/releases

* Switchboard[5]: 2.0.0, 2.1.0, 2.2.0 and 2.3.0.

Most of the integrations planned initially have been achieved. Both the VLO and VCR have added *mopinion* to collect user satisfaction feedback for the VA reporting (this does not apply to the Switchboard). The VLO has been integrated with both the VCR and the Switchboard and the VLO integration with B2FIND has been further improved. The VCR has extended the submission endpoint for third party integrations to support the submission of more metadata so that a more complete collection can be created. This has been integrated with the VLO. The Switchboard is integrated with both the VLO and the VCR. The B2DROP plugin, integrating B2DROP with the Switchboard, has been improved and updated for the latest release of B2DROP. In addition to this plugin a format specification[6] and API[7] have been developed. This allows tool providers to better describe the nature and method of interaction to the switchboard. The different options to integrate with the Switchboard as a resource provider have also been formalized[8].

Both the VCR and the Switchboard were initially planned to be integrated with B2SHARE. This has been discussed with the B2SHARE team roughly halfway of the project. Due to internal priorities this is still on the roadmap for the B2SHARE team.

We have requested resources from EOSC-hub to deploy a VLO instance and to experiment with various improvements for our thematic services in general, such as log analysis. For the VLO instances we specifically require high Input/output operations per second (IOPS). We received offers from three providers: CESGA, ReCaS and CESNET. After evaluating the providers it turned out that latency to Spain (CESGA) and Italy (ReCaS) was too high to run our production load and after discussing this with the providers we concluded there was no room for significant improvements. Therefore we continued to use the resources provided by CESGA and ReCas to deploy and test development versions of our thematic services. The CESNET instance is running our elasticsearch / kibana stack and is currently processing logs of all our thematic service instances, including production.

In addition to the planned activities related the thematic services, we have also used EOSC-hub resources to help us run the Reprolang 2020 workshop[9].

2.4 Impact and exploitation

The usage is measured using the following Virtual Access metrics, as reported to WP13. We left out the first very short reporting period (M7-M8) because it overlaps with the summer vacation and is too short to be representative. For full details we refer to the deliverable on Virtual Access.

Overall we can conclude from the numbers presented below that there is a clear and continuous increase of the indicators used to measure the degree of use and connectivity of the CLARIN thematic services.

[5] https://github.com/clarin-eric/switchboard/releases
[6] https://github.com/clarin-eric/switchboard-doc/blob/master/documentation/ToolDescriptionSpec.md
[7] https://github.com/clarin-eric/switchboard-doc/blob/master/documentation/ToolCallAPI.md
[8] https://github.com/clarin-eric/switchboard-doc/blob/master/documentation/IntegrationProvider.md
[9] https://lrec2020.lrec-conf.org/en/reprolang2020/

Number of visits to metadata search portal

Measurement definition: number of registrations, reported is number of visits over a certain timespan – measured using Matomo

	Baseline 2017	**Period 2** M9-M17	**Period 3** M18 - M27	**Period 4** M28 - M36
Visits / Month	425	528	647	1013

Number of harvested metadata records

Measurement definition: the number of metadata records harvested via OAI-PMH and inserted into the Virtual Language Observatory.

	Baseline 2017	**Period 2** M9-M17	**Period 3** M18 - M27	**Period 4** M28 - End
Records	896473	909388	907429	1203949

Number of virtual collections registered

Measurement definition: the number of virtual collections made publicly available via the virtual collection registry over a certain timespan.

Note: the baseline is low since not much publicity was made before the EOSC-hub release.

	Baseline 2017	**Period 2** M9-M17	**Period 3** M18 - M27	**Period 4** M28 - End
Collections	0	7	7 (and 3 tests)	13

Number of connected processing tools via the LR Switchboard

Measurement definition: the number of web applications registered at the Language Resource Switchboard that can process incoming requests.

	Baseline 1 January 2018	**Period 2** M9-M17	**Period 3** M18 - M27	**Period 4** M28 - End
Tools	60	70	72	154

Number and names of the countries reached

Measurement definition: measurement for metadata search portal, based on IPs, measured using Matomo

	Baseline 2017	**Period 2** M9-M17	**Period 3** M18 - M27	**Period 4** M28 - End
Countries	89	101	112	155

Satisfaction

Measurement definition: 5-point scale Customer Satisfaction measurement, measured using Mopinion. Mopinion was integrated into the Virtual Language Observatory in July 2018.

	Baseline 2017	**Period 2** M9-M17	**Period 3** M18 - M27	**Period 4** M28 - End
Customer Satisfaction score	not applicable	3.9 (120 responses)	3.8 (105 responses)	3.9 (74 responses)

2.5 Lesson learnt

For some of the integrations we have defined with external services, we would like the external service to implement some additional functionality in order to realise such integration. We realised this is a risk since this makes us dependent on the available time and priorities of these external teams. Therefore, we designed the integration work to be as minimal as possible. For the Switchboard and B2DROP integration, we have developed the nextcloud plugin and the B2DROP team only has to enable and test this plugin. We also take care of the updating of this plugin. With the Switchboard B2SHARE and VCR B2SHARE integrations, there was no plugin framework available, so we started discussions with the B2SHARE team to get this effort on their roadmap. While this succeeded, the work got pushed back repeatedly. This shows that this type of integration is a risk you have to properly account for.

All three thematic services have been integrated with the EOSC-hub marketplace. For the VLO we planned a specific task to support requests for community specific deployments. Unfortunately, there have been no such requests via the EOSC-hub marketplace so far. All three services are open access and this model probably does not match very well with the marketplace model centring around the “checkout” of a service which was the marketplace model when the marketplace was first released. Currently the marketplace model has improved a lot and now also supports (fully) open access services. This is a better fit for our thematic services.

2.6 Future plans beyond EOSC-hub

The thematic services will be further supported and developed by CLARIN ERIC. Additional improvements will be ensured through CLARIN's participation in the SSHOC project[10] and in the EOSC Future project. For each of the services we are working on roadmaps for the future. The following paragraphs give an idea of the things we are considering.

[10] https://sshopencloud.eu/

For the VLO we envision the implementation of a publicly available API that provides access to the records of the VLO, enabling integrations that build on the processing carried out by and for the VLO, and thus opening up possibilities for extended functionality serving the needs of scholars across domains, as well as enhanced end user experience.

For the VCR we envision the implementation of versioning of collections and adding support for collaborative editing. These requirements have been collected within the SSHOC project and are driven by use-cases from SSHOC communities.

For the Switchboard we envision the implementation of support to input multiple files in the same request. This will enable further integrations with tools that operate over complementary file sets (e.g. EAF + MP4), tools that operate batches of files or collections and data providers that can send multiple files to the Switchboard at once.

3 DODAS

3.1 Service description

Service/Tool name	**DODAS**
Service/Tool url	http://dodas-iam.cloud.cnaf.infn.it/
Service/Tool information page	https://dodas-ts.github.io/dodas-apps/
Description	DODAS is a cloud enabler for scientists seeking to easily exploit distributed and heterogeneous clouds to process, manipulate or generate data.
Value proposition	Lower technical barriers to cloud adoption
Customer of the service/tool	Researcher, Resource Provider, Research Communities
User of the service/tool	Same as above
User Documentation	https://dodas-ts.github.io/dodas-apps
Technical Documentation	https://dodas-ts.github.io/dodas-apps/#developers-guide
Product team	INFN
License	Apache License Version 2.0
Source code	https://github.com/dodas-ts
Testing	-

DODAS is a Platform as a Service whose aim is to guarantee deployment of complex and intricate setup on “any cloud provider” with almost zero effort. As such it implements the paradigm of Infrastructure as code: driven by a templating engine to specify high-level requirements. DODAS allows instantiating on-demand container-based clusters to execute software applications.

DODAS completely automates the process of provisioning by creating, managing, and accessing a pool of heterogeneous computing and storage resources. As a consequence, it drastically reduces the learning curve as well as the operational cost of managing community-specific services running on distributed clouds.

DODAS allows instantiating on-demand complex infrastructures over any cloud with almost zero effort and with very limited knowledge of the underlying technical details. In particular DODAS provides the end user with all the support to deploy from scratch a variety of solutions dedicated (but not limited) to scientific data analysis. For instance, with pre-compiled templates the users can create a K8s cluster and deploy on top of it their preferred Helm charts all in one step. DODAS provides three principal baselines ready to be used and to be possibly extended:

- an HTCondor batch system
- a Spark+Jupyter cluster for interactive and big-data analysis
- a Caching on demand system based on XRootD

3.2 Initial ambition (in 2018)

Since the beginning of the project, the aim of DODAS service has been the provision a user-friendly solution for science communities in order to exploit opportunistic computing, intended as resources not necessarily or permanently dedicated to a specific experiment and/or activity; elastic extension of existing facilities, to absorb peaks of resource usage; generation of on-demand batch systems and/or Machine/Deep Learning facilities for data processing.

The initial objective was to target:

- User/researcher who needs to exploit opportunistic computing. An effective usage of compute and data resources, for medium to large data processing requires quite advanced IT skills. DODAS, by providing automation, abstraction and self-healing capabilities represents an ideal solution.
- Site manager who needs to elastically extend existing facilities. Sites and facilities already support communities and experiments for any related computing activity.
- Researchers who need a data analytics Infrastructure as a Service. New approaches to data analysis are more and more required for user communities. A frequent scenario is that users need to have access to facilities where to develop, test and train models. Not only, but also some facilities where to perform inference as well.

To tackle the above concepts, the specific actions planned EOSC-hub were to extend the experience previously done and to consolidate:

- Enhancing AAI integration,
- Monitoring account capabilities,
- Further integration with Data management,
- Integration with INDIGO-Paas Orchestration.

Since the start of the project a solid exploitation program was established in order to get scientific communities onboarded with school and training programmes finalized to disseminate the DODAS project including hackathons, seminars, PhD schools, and training courses.

3.3 Final software architecture and integration

The DODAS service is meant to be highly customizable to be adapted to the evolving needs of scientific communities. From a technical point of view, DODAS relies on a service composition model providing a highly versatile portfolio to the adopter. There are four main pillars in the DODAS architecture which can be summarized as:

- abstraction: both in terms of software application and dependency description, and of underlying IaaS level cloud infrastructures;
- automation: it refers to software and application setup in order to manage resources and orchestrate software applications;
- multi-cloud support: to deal with multiple heterogeneous Cloud infrastructures;
- flexible AAI: authentication, authorization, delegation, and credential translation primitives providing a secure composition of the various services participating in the DODAS workflow.

Figure 3.1 shows these pillars. They have been realised mostly using services also available in the EOSC-hub portfolio: Identity and Access Management (INDIGO-IAM) and Token Translation Service (TTS), PaaS Orchestrator and Infrastructure Manager (IM). The composition of these services represents the so-called PaaS Core Service of DODAS.

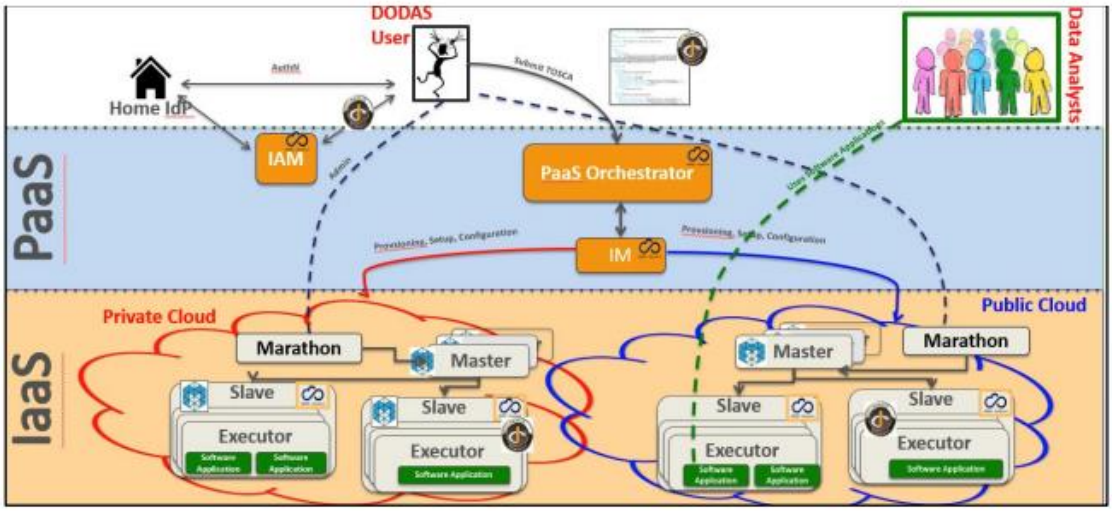


Figure 3.1: High level schema of the DODAS architecture

The DODAS architecture integrates since the beginning several of the EOSC-hub solutions such as Infrastructure Manager, Identity and Access Management (IAM) and Token Translation service.

During the project, DODAS also integrated OneData as a possible solution for the data ingestion and for transparent data access. CVMFS stratum 0 and 1 have been also integrated in order to support newer use cases coming from AMS-02 experiment requirements.

In the final software architecture several components have been further enhanced and integrated. Particularly:

- Monitoring: A key to the usability of the system is of course also the monitoring system. As DODAS is a very customizable and versatile service in terms of computing features provided. The original monitoring system has been evolved towards a Prometheus solution supporting custom exporters. A Grafana dashboard is also deployed automatically.
- Storage and data services: Data management support has been further extended to support a cache-based mechanism which includes the creation of automatic procedures to enable a

fast and effortless deployment of cache clusters in a cloud environment. With respect to the original Xrootd and OneData technologies, the support has been further extended to generic S3 cloud storage. The target in this respect has been MinIO.

- Integration with INDIGO PaaS Orchestrator: Particular effort has been spent in order to improve the integration with the user-friendly dashboard that can be used to submit deployment requests in a simple and straightforward way. The dashboard hides the complexity of the TOSCA templates providing simple forms for the deployment configuration and customization. The user can retrieve the list of her/his deployments with details like the creation time, the status, the provider hosting the deployment resources and logs.
- Resource auto scaling: a lot of investment has been put in order to define a generic architectural approach to the autoscaling challenge. The main point here has been the development of a Prometheus based approach which allows to define any custom metric to be exported and thus used by the cluster orchestrator to decide if to enlarge or shrink the deployed services.

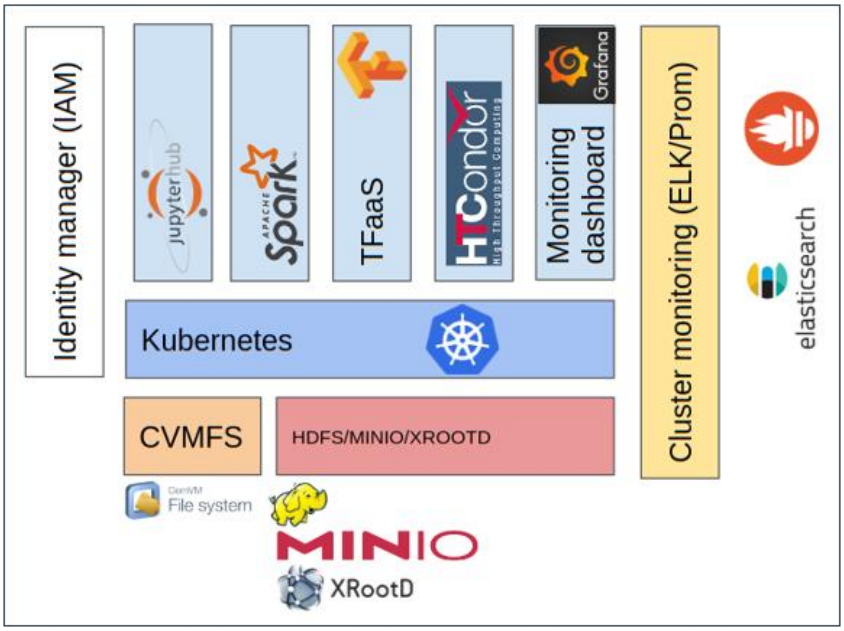


Figure 3.2: High level view of the end user services provided by DODAS to the user. Any part of the software stack can be composed "lego style" on demand.

For what concerns the end user perspectives, the service and functionalities and solutions provided by DODAS, the final software includes a variety of lego composable software's as shown in Figure 3.2.

Finally, DODAS stable version is published via EOSC both on the Marketplace and on the Service Catalogue.

3.4 Impact and exploitation

The impact of the DODAS Thematic Service can clearly be seen from several perspectives. First of all looking at several Virtual Access metrics that have been defined and reported to WP13 for each reporting period (summarised in the tables below), it is possible to see the increase during the project.

	Period 1 M3-M8	**Period 2** M9-M17	**Period 3** M18 - M27	**Period 4** M28 - End
# Jobs	30k 11894 CMS Experiment at CERN + 17465 CMS OpenData Project at LHC + ~20k AMS02 experiment operating on the International Space Station (ISS)[11].	~500k CMS ~600k AMS	~200k CMS ~ 150k FERMI ~ 200k AMS	~150k CMS ~40k FERMI ~645k AMS ~8k Theoretical physicist

	Period 1 M3-M8	**Period 2** M9-M17	**Period 3** M18 - M27	**Period 4** M28 - End
#Deployments	622	1084	647	646

	Period 1 M3-M8	**Period 2** M9-M17	**Period 3** M18 - M27	**Period 4** M28 - End
Scientific communities	CMS, AMS, ImpCollege, OpenData	CMS, AMS, Imperial College, OpenData and Virgo	CMS, AMS, Virgo, Fermi	CMS, AMS, Virgo, Fermi and Theoretical Physicist

	Period 1 M3-M8	**Period 2** M9-M17	**Period 3** M18 - M27	**Period 4** M28 - End
Providers	Imperial College + T-System	Imperial College, T-System and Google Cloud and Amazon	ASI, EGI Federated Clouds, AWS	ASI, AWS, INFN-Cloud

[11] https://cms.cern/org/cms-scientific-results; https://cds.cern.ch/record/2637774/files/felipeCordero.pdf; https://ams02.space/

	Period 1 M3-M8	**Period 2** M9-M17	**Period 3** M18 - M27	**Period 4** M28 - End
Countries	Italy	Italy, UK and Switzerland	Italy, UK and Switzerland, USA	Italy, Switzerland, China, USA

Figure 3.3 shows the trends over the years of most of them demonstrating the impact. The number of newly registered users per reporting period has steadily increased during the project.

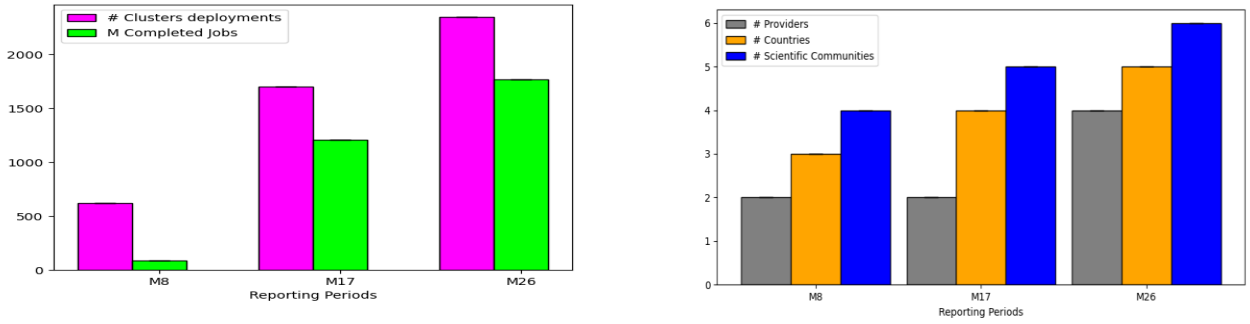


Figure 3.3: Trends of Virtual Access metrics over the reporting periods of the project: Jobs and Cluster deployments on the left while the plots on the right show the number of reached providers, countries and Scientific communities.

From another perspective, the picture in Figure 3.4 shows not only the major achievements from the scientific communities and the resources providers point of view but also highlights the connections of DODAS with other EU funded projects in the field such as HelixNebula Science Cloud and ESCAPE.

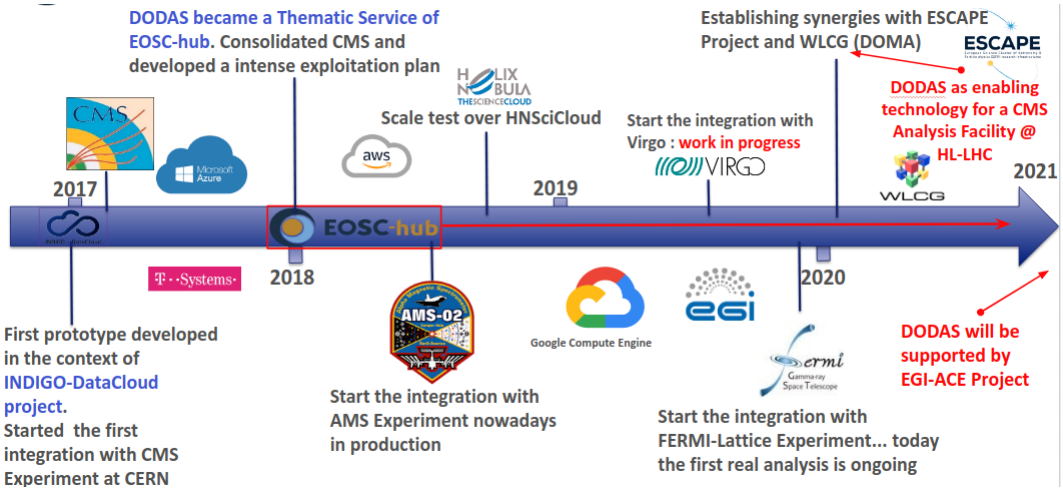


Figure 3.4: The picture shows the timeline of the DODAS Thematic Services reporting the major achievements in terms of communities, service providers and connection with external EU projects.

During the EOSC-hub project DODAS has been used to integrate and exploit also the EGI cloud compute resources. In this respect DODAS has been used as a PaaS federation layer on top of the EGI Federated Cloud.

Finally, the very successful integration of two major communities (AMS02 and FERMI) proves the impact of DODAS. Scientific data analysis has been done using DODAS provided resources and, the successful approach led some of the members to propose a DODAS approach also for future experiments such as HERD.

3.5 Lesson learnt

The planned integrations activities defined at the beginning of the project have been mostly achieved (only the accounting integration for dynamic cluster was not completed). As previously stated, the DODAS thematic service has been successfully integrated into the EOSC marketplace and the EGI cloud infrastructure.

A few lessons learnt while integrating and operating the DODAS Thematic Service are worth to be summarized with the aim to provide useful feedback for further improve the process of the EOSC building:

- The majority of users of DODAS find the service not via the EOSC Marketplace. The modus-operandi of main DODAS communities and the fact that DODAS resources are provided on a free basis to users might not help the adoption of a marketplace-based solution.
- Enhancing the strategy towards an integrated/federated AAI may be useful to facilitate and to improve both the user experience and the service integration process. Beside the huge effort put in this topic, the current status seems still suboptimal.
- Storage and data management is an area where to keep the investments going on considering the big trend in the evolving and development of Data Lake models for multiple disciplines.
- Training activities: A very huge amount of effort should be invested in this respect in order to help community toward the paradigm shifts on:
    - resources/e-infrastructures exploitation,
    - software management in a native cloud environment,
    - data access and data sharing.
- The training program plays a key role for what regards the user engagement. The DODAS experience teaches that the majority of new communities and users have been on-boarded thanks to the training activities.
    - training should be also very technical and specific, oriented to application porting.
- Integration cannot be seen as activity performed once. Even the more mature communities need to be supported continuously to adapt to the fast changes in the evolving technologies on the one hand, and to integrate the requirements arising from a developing and dynamic environment on the other hand.

3.6 Future plans beyond EOSC-hub

The DODAS Thematic Service will keep operating under the EGI-ACE project in order to deliver compute and data analysis capabilities to the scientific communities. Moreover, DODAS will continue its support to the adopter in the context of the INFN-Cloud National project.

The strong connections of the DODAS Thematic Service in the context of several working groups of WLCG such as DOMA will allow gathering key feedbacks to evolve the service in a coherent way with respect to the future computing models of many HEP and Astroparticle experiments.

From a technical perspective, the plans beyond the EOSC-hub will be toward the direction of improving the support of interactive or quasi-interactive data analysis. This roughly translates into the needs of improving the exploitation of specialized hardware such as GPU FPGA, NVMe, etc. These are mandatory steps to satisfy the main request of having platforms for a high throughput data analysis.

4 ECAS

4.1 Service Description

Service/Tool name	**ENES Climate Analytics Service (ECAS)**
Service/Tool url	CMCC ENDPOINT: https://ecaslab.cmcc.it/web/home.html DKRZ ENDPOINT: https://ecaslab.dkrz.de/home.html
Service/Tool information page	https://www.eosc-hub.eu/services/ENES%20Climate%20Analytics%20Service https://portal.enes.org/data/data-metadata-service/climate-analytics-service
Description	The ENES Climate Analytics Service is a server-side processing service which offers a virtual work environment based on Jupyter notebooks, allowing users to process and analyse data from multiple disciplines using Python. Support for fast computations is provided via the Ophidia HPDA framework.
Value proposition	ECAS enables scientific end-users to perform data analysis experiments on large volumes of multidimensional data by exploiting a server-side, PID-enabled, and parallel approach and aiming to improve reusability of data and workflows (FAIR approach).
Customer of the service/tool	Climate modelling community; direct downstream usage communities
User of the service/tool	Researchers; Scientific users
User Documentation	https://ee-docs.readthedocs.io/en/latest/ https://docs.egi.eu/users/cloud-compute/ec3/apps/ecas/ http://ophidia.cmcc.it/documentation/index.html https://ecaslab.dkrz.de/home.html
Technical Documentation	https://github.com/ECAS-Lab/ecas-authentication https://github.com/ECAS-Lab/ecas-accounting https://github.com/ECAS-Lab/ecas-monitoring http://ophidia.cmcc.it/documentation/admin/index.html
Product team	Fondazione Centro Euro-Mediterraneo sui Cambiamenti Climatici (Fondazione CMCC); Deutsches Klimarechenzentrum GmbH (DKRZ)
License	The Ophidia code is available on GitHub under GPLv3 license; additional components for ECAS (Docker workflow components) available under BSD license; ECAS-B2SHARE Python client available under MIT license.
Source code	https://github.com/ECAS-Lab; https://github.com/OphidiaBigData
Testing	ECAS single-instance VMI uploaded to the EGI AppDB: https://appdb.egi.eu/store/vappliance/ecas

	Deployment test using the EGI AppDB VMops dashboard: https://dashboard.appdb.egi.eu/vmops Multi-node ECAS environment, dynamically provisioned on the EGI FedCloud through the EC3 LToS service: https://servproject.i3m.upv.es/ec3-ltos/index.php

The ENES Climate Analytics Service (ECAS) enables scientific end-users to perform data analysis experiments on large volumes of multidimensional data (e.g. NetCDF data format), by exploiting a PID-enabled, server-side, and parallel approach. The service is aimed at providing a paradigm shift for the ENES community with a strong focus on data intensive analysis, provenance management, and server-side approaches as opposed to the current ones that are mostly client-based, sequential and with limited/missing end-to-end analytics workflow/provenance capabilities. ECAS consists of multiple integrated components, centred around Big Data Analytics frameworks, initially Ophidia[12]. ECAS has been integrated with B2DROP, ESGF, IAM, EGI Check-in, Onedata (DataHub), EGI FedCloud, JupyterHub, and the ECAS-Lab web portal.

The technical details are further described in the previous deliverables [R1], [R2].

4.2 Initial ambition (in 2018)

At the beginning of the project, a dedicated goal of the ECAS service was to engage end-users directly, interact with them and induce a cultural change for the scientific workflow.

ECAS aimed at a wide range of users from several communities:

1. Scientific users from the core climate modelling community, who did not have access to sufficient computing resources for large climate data analysis experiments;
2. Scientific users from the direct downstream usage communities, such as climate impact studies, who were not familiar with the data design and ESGF data distribution mechanics and, therefore, wanted to benefit both from easier accessibility of processing and transparent selection of input data without needing to understand data locality or arrange data transfers;
3. Scientific users from other research communities, who deal with the same data model (multidimensional data) and similar challenges (large scale data analysis).

Based on the experience coming from INDIGO-DataCloud, additional communities that could benefit from the ENES Climate Analytics Service are: (i) EMSO (European Multidisciplinary Seafloor and water-column Observatory), (ii) LBT (Large Binocular Telescope) and also (iii) LifeWatch. In the first phase of the project, ECAS aimed to address the ENES-related use cases (points 1 and 2).

At the same time, the activities would bring the ENES data infrastructure (IS-ENES, the European contribution to the Earth System Grid Federation) closer to an amalgamation with EUDAT, EGI and INDIGO.

[12] http://ophidia.cmcc.it/

In such a context, the defined roadmap included:

- several activities to fully integrate ECAS into the EOSC-hub service portfolio;
- iterative releases according to the project deadlines and planning;
- user engagement, through several dissemination activities and training events;
- integration into the EOSC marketplace and the cloud-based resources provided by EGI (EGI FedCloud);
- contribution to Open Science/Research in terms of sharing and re-use of workflows, collaborative experiments, publication of results.

4.3 Final software architecture and integration

The following figure provides an overview of the final ECAS architecture with all the integrated services.

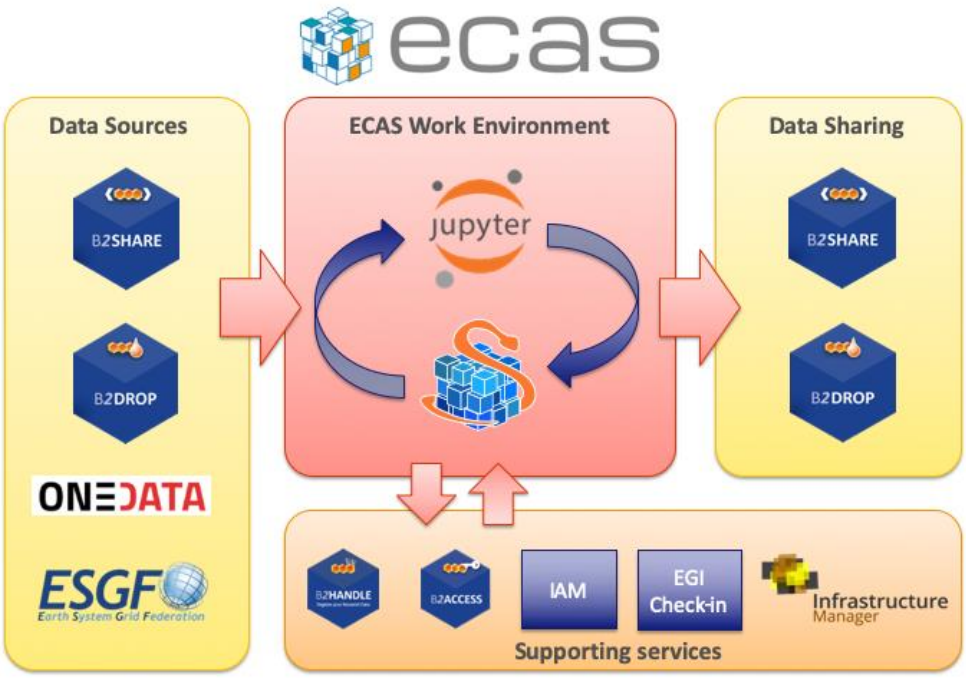


Figure 4.1: ECAS software architecture

The Ophidia framework provides the key scalable, parallel analytics capabilities. To enable easy data provisioning, it is integrated with B2DROP, Onedata and ESGF. The integration with JupyterHub, IAM and EGI Check-in enables users to easily access the service and re-use existing Python scripts as well as modify them or create entirely new ones in a fully server-side approach. More details are reported in the EOSC-hub Deliverable D7.1 [R1].

ECAS consists of two main instances hosted by CMCC and DKRZ, respectively. The current configuration of the available resources is the following:

- ECAS@CMCC:
    - 1 *client node* with 4 cores, 8GB memory, 12GB disk, hosting the client-side components (e.g. the Ophidia CLI and PyOphidia module) and the front-end services (e.g. JupyterHub);
    - 1 *server node* with 20 cores, 256GB memory, 4,7TB disk, hosting the Ophidia Server and complementary administration, monitoring and accounting services;
    - 5 *compute nodes*, 20 cores/node, RAM 256GB/node, 60TB disk space shared via GFS plus local disks, dedicated to the Ophidia framework and the I/O servers.
- ECAS@DKRZ:

Instance with Ophidia backend:

- 1 *client node* (VM) with 4 cores, 15GB memory, 50GB local disk + 10TB NFS for the JupyterHub and Docker services;
- 1 *server node* (VM) with 4 cores, 4GB memory, 30GB disk, hosting the Ophidia server;
- 1 *compute node* (Blade) with 40 cores, RAM 256GB, 2TB local disk, executing Ophidia I/O server + 10TB NFS disk space.
- CMIP5 and CMIP6 (read-only) *data pool*.

New instance with xarray, dask backend (see “future plans” and “lessons learned” sections):

- dynamic number of client nodes (dynamically allocated on HPC system)
- dynamic integration of user defined jupyter kernels possible
- dynamic number of compute nodes (dynamically allocated on HPC system)
- 5 PByte CMIP (read-only) climate data pool
- intake based data catalog

Most of the activities concerning the integration of B2DROP and Onedata into the ECAS environment for data sharing, alongside IAM as the AAI solution and ESGF data sources, started during the first year of the project. Additional actions undertaken during the first year regarded the software containerization of the ECAS environment, some improvements to the usability of the data analytics features from JupyterHub, the integration with the EOSC-hub accounting and monitoring, documentation, as well as the implementation of several demonstration Jupyter notebooks, and extensive training. Integration activities for the first year of the project have already been presented in detail in deliverable D7.2 [R2].

During the second year of the project, several of these integration activities were finalized. In particular, they concerned a stronger integration of B2DROP and Onedata data services, the full integration of IAM for user authN/authZ (from JupyterHub to the data analytics services) and some extensions to the core services for accounting and monitoring purposes. Moreover, the Ophidia framework and its Python interface have been extended to further improve the user experience and

allow the development of other climate-oriented use cases; additional Jupyter notebooks have also been implemented as both demonstrators of the environment feature and self-study/training material. The views for user accounting have been fully implemented and the ECAS environment has been successfully integrated into the EOSC Marketplace[13] and the EGI Federated Cloud deployment services[14].

Furthermore, additional extensions were carried out during the third year of the project, targeting a stronger integration of ECAS (through PyOphidia) with the Python ecosystem and Jupyter Notebooks as well as the integration of ECAS with EGI Check-in. More details about the above mentioned activities are reported in the EOSC-Hub deliverable D7.4 [R7].

Following the amendments in the second part of the project, additional activities have been carried out to further strengthen the integration with EOSC-hub services. In addition, some bug fixing, and minor improvements have been performed based on users' feedback and requests to improve the robustness of the ECAS service. Moreover, in order to enhance the overall system stability, some of the components of the ECAS instance deployed at CMCC have been migrated on a new and more powerful server node equipped with greater hardware resources (20 cores, 256GB memory). Among them:

- the *Ophidia server*
- *Grafana* and the underlying *Influxdb* database, both installed and run as Docker containers.

All developments, integration efforts and training material concerning ECAS are stored on a dedicated GitHub repository[15], while extensions and adaptations to other related tools have been published on the corresponding GitHub repositories (i.e. for the Ophidia framework[16]).

The next subsections provide more details about the integration activities finalized in the last part of the project.

Integration of ECASLab with Onedata

Concerning the integration of ECASLab with *Onedata*, both the *Oneprovider* instance, deployed at the CMCC SuperComputing Center to support the ECAS_space with actual storage resources, and the *Oneclient* service, used to mount the ECAS_space on the cluster nodes, have been upgraded to the latest versions to make them compliant with the associated *Onezone* (https://datahub.egi.eu/).

[13] https://marketplace.eosc-portal.eu/services/enes-climate-analytics-service
[14] https://www.egi.eu/about/newsletters/elastic-deployment-of-ecas-on-egi/
[15] https://github.com/ECAS-Lab/
[16] https://github.com/OphidiaBigData

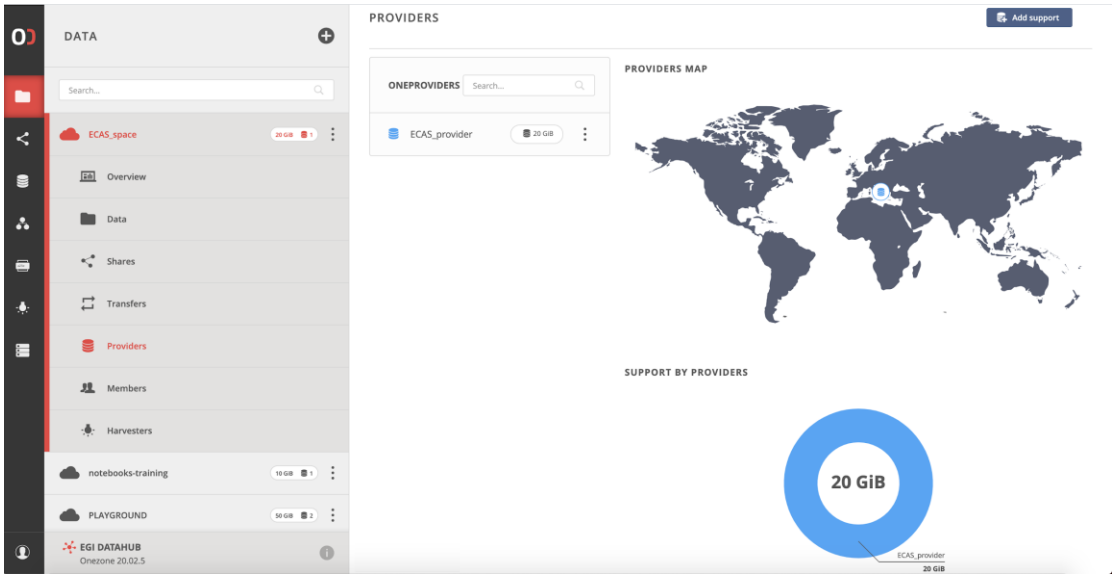


Figure 4.2: ECAS_space supported by CMCC from the Onezone web interface.

The ECAS_space is used (as a proof-of-concept) to store some NetCDF files in order to provide ECAS users with a read-only access to a sample data repository hosted in a Onedata space allowing analysis on this shared data (Fig. 4.3).

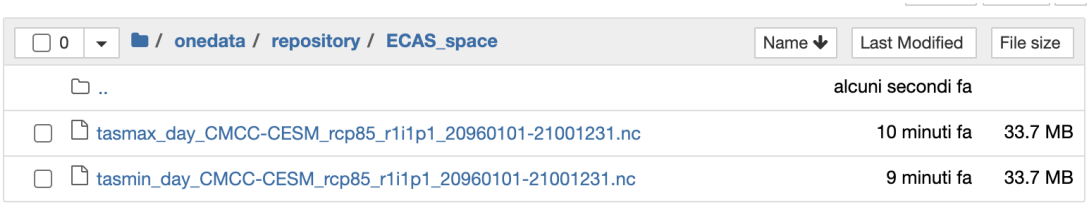
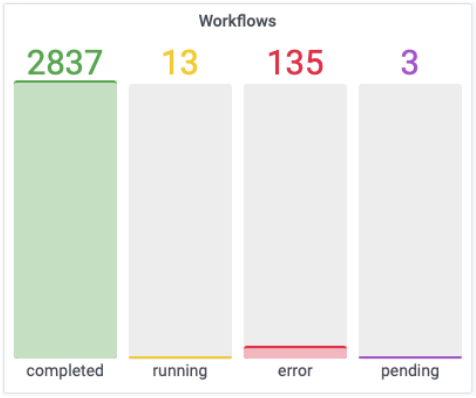


Figure 4.3: Repository

ECASLab monitoring system

As mentioned above, the Grafana-based monitoring system deployed on the ECASLab instance at CMCC has been migrated on a more powerful node as Docker container services, but also extended and revised both at system and application level, to improve monitoring of the ECAS environment and resource usage. The available dashboards have been enhanced to better identify which operator/workflow is being executed and its current execution status, and to provide aggregated information over time.

More specifically, a panel that shows the total number of all running, failed, pending, waiting and completed workflows has been integrated in the *Application Metrics* dashboard.

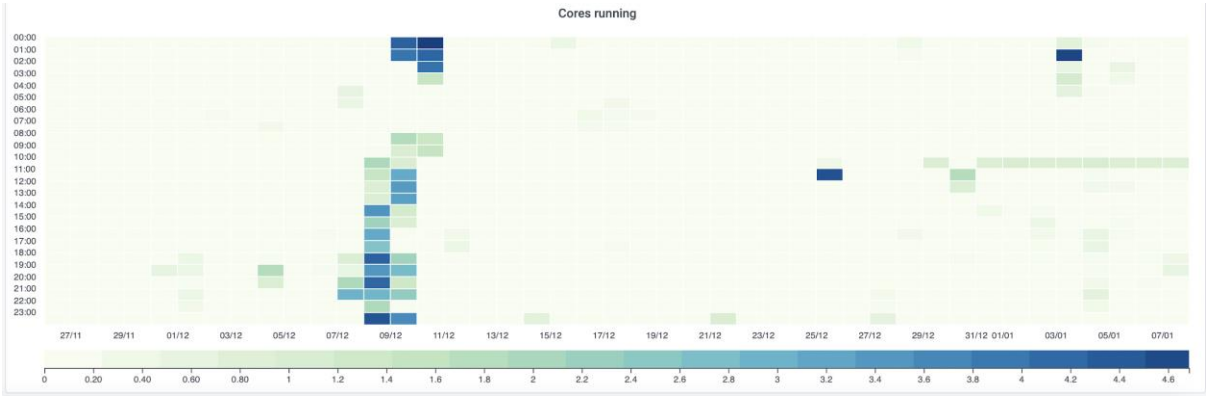


The workflow progress table has also been modified in the same dashboard to introduce additional information and features such as sorting and filtering the information provided for each column (e.g., timestamp, workflow name, final status, etc.).

Workflow Progress

Time	Workflow	Completed Task	Total Task	Progress Ratio		Status
2021-01-20 14:36:32	Basic workflow #143	3	3	100%		ERROR!
2021-01-20 14:35:55	Basic workflow #142	3	3	100%		ERROR!
2021-01-20 14:29:05	oph_list #139	1	1	100%		completed!
2021-01-20 14:28:31	Basic workflow #138	0	3	18%		running ...
2021-01-20 11:57:59	Basic workflow #134	1	3	33%		ERROR!
2021-01-20 11:16:48	Basic workflow #132	3	3	100%		completed!
2021-01-20 11:10:59	Basic workflow #129	1	3	33%		ERROR!
2021-01-20 11:09:01	Basic workflow #128	1	3	33%		ERROR!
2021-01-20 11:05:40	Basic workflow #124	1	3	33%		ERROR!
2021-01-19 18:25:57	oph_list #123	1	1	100%		completed!
2021-01-19 18:22:18	Loop operations #122	0	15	0%		pending ...
2021-01-19 18:19:03	Basic workflow #121	1	3	33%		ERROR!
2021-01-19 18:18:43	oph_list #120	1	1	100%		completed!
2021-01-19 18:14:52	Basic workflow #119	1	3	33%		ERROR!

In addition, the graph that provided the instantaneous number of cores running over time has been replaced with an hourly heatmap showing the hourly weighted average of the number of running cores. This new view provides the graphical means to track the daily load distribution more efficiently on the system.



As for the *Infrastructure Metrics* dashboard, two additional panels showing the percentage of main and swap memory used on each node have been inserted. This new visualization allows an easier identification of high workload usage events.

Moreover, the infrastructure view and the corresponding monitoring scripts have been extended to also provide information about the disk space on each node and the disk space from the shared file system at the ECAS@CMCC cluster

The dashboard has been further expanded with two additional sections showing the monitoring metrics of the client and front-end nodes, in addition to the already monitored compute nodes.

Integration of ECAS into the EGI Federated Cloud

The integration of ECAS into the EGI deployment services has also been improved. Specifically, two different scenarios were considered for the integration, as already described in previous deliverables:

- new versions of the ECAS Virtual Appliance providing a ready-to-use ECAS environment have been uploaded to the EGI Applications Database[17] to include the various integration activities undertaken in the last project period, with the final one including the full set of integration described in this deliverable, e.g., Grafana-based monitoring, OneData support, etc.
- the Ansible Role[18] has also been updated and extended to integrate these new features into the elastic deployment of ECAS on the EGI FedCloud through the EC3 LToS service[19].

In the latter case, the integration required some adaptations in order to support automatic configurations of the services for an elastic multi-node ECAS deployment.

In particular, concerning Grafana-based monitoring, the two dashboards used to monitor the ECAS cluster (at system and application level) have been adapted with respect to the ECASLab versions in order to reflect the elastic behaviour of the cluster in terms of number of active working node instances, which may vary according to the actual user workload.

Figure 4.4 and Figure 4.5 depict the sections of the *Infrastructure Metrics* dashboard showing the monitoring metrics of the front-end node (named *oph-server*) and of the active working nodes (only one in this example, named *oph-io1*), respectively.

[17] https://appdb.egi.eu/store/vappliance/ecas

[18] https://github.com/OphidiaBigData/ansible-role-ophidia-cluster

[19] https://servproject.i3m.upv.es/ec3-ltos/index.php

Figure 4.4: Infrastructure Metrics dashboard showing the monitoring metrics of the front-end node

Figure 4.5: Infrastructure Metrics dashboard showing the monitoring metrics of the active working node

Figure 4.6: a section of the Application Metrics dashboard and, in particular, the workflow progress table and a panel visualizing information about the total number of workflows, grouped by status.

The corresponding EGI documentation page[20] has been extended accordingly to include the new developments. In addition, a new GitHub repository named *ECAS-monitoring*[21] has been created under the *ECAS-Lab* project repo[22], to store the implemented monitoring scripts and the Grafana JSON dashboards used as templates by the Ansible role during playbook execution.

The updated version of the Ansible-based deployment also integrates the Onedata ECAS_space presented above through the Onedata client tool, that was used to mount the repository hosting climate data of interest for the community. Besides the integration of the components into the deployment, the Ansible role has been updated to also include the latest version of the Ophidia framework, the python modules and the training notebooks. Finally, a new set of Python libraries have been included to enrich the Python ecosystem.

[20] https://docs.egi.eu/users/cloud-compute/ec3/apps/ecas/

[21] https://github.com/ECAS-Lab/ecas-monitoring

[22] https://github.com/ECAS-Lab/

User experience

To support user training and demonstration activities, new Jupyter Notebooks integrating some of the key capabilities of ECAS and showing how to exploit the main features offered by the integrated tools (e.g. Onedata, B2DROP, etc.) have been implemented and updated and made available on the ECASLab GitHub repository[23], as well as directly through the ECAS service for registered users.

Moreover, the CMCC ECAS website has been updated to improve the graphical user experience and to extend the contents. Some sections have been redesigned to include the new development activities, while some contents have been modified according to the improvements carried out during the final period of the project.

Data availability and cataloguing

The data collections available via the ECAS service at CMCC have been extended to include additional data collections of interest to the climate community. In particular, new CMIP6 datasets related to the precipitation and air temperature variables at different time frequencies have been downloaded from the ESGF data catalog for a total of 7 TB. They have been properly organized in the local archive and made available to registered users via the shared file system.

To make the data search and discovery workflow easier for ECAS users, a catalog file containing the data asset locations and the associated metadata has been created as an underlying database for the *intake-esm*[24] *Python* tool. In this way, users can easily index and access the climate datasets available on ECAS to perform their own analysis on the selected data directly from the Jupyter environment, without having to browse for the files on the file system.

Further pre-processing procedures have been applied to improve and streamline all analysis activities on such data.

Metadata management and climate data FAIRness

Moreover, a feasibility study has been carried out to also deal with metadata management and explore the possibility of describing and enriching data by means of any attribute and aiming at FAIR principles for scientific data management.

In this regard, the integration of ECASLab with Onedata represents a good starting point, as Onedata comes with extensive support for metadata management[25], which can be used to describe all kinds of resources including files, directories, spaces and users. Indeed, Onedata allows assigning custom key-value pairs as well as arbitrary metadata documents to each resource in one of the supported metadata formats (JSON, RDF).

In addition, CMIP* data outputs follow the standard NetCDF Climate and Forecast (CF) Metadata convention and specifically adhere to the standard variable names, units, dimensions, axis, required 'coordinates' attribute, bounds and stored direction for overall intercompatibility. All these

[23] https://github.com/ECAS-Lab/ecas-notebooks

[24] https://github.com/intake/intake-esm

[25] https://onedata.org/#/home/documentation/stable/doc/using_onedata/metadata.html

components are defined in the Data Reference Syntax (DRS) documents[26][27], which provide a clear and structured set of conventions to facilitate the naming of data entities within the data archive and the naming of files delivered to users. In order to facilitate documentation and discovery, the DRS employs Controlled Vocabularies (CVs) that are useful to develop category-based data discovery services.

In such a context, this well-structured metadata schema provided for CMIP* data could be integrated with the features offered by the Onedata platform in terms of metadata management; this would help define proper publishing procedures and build a robust workflow which, in turn, would enable easier searching and indexing of climate datasets in ECAS, thus increasing FAIRness of climate data in EOSC.

4.4 Impact and exploitation

In order to detail the access activity with virtual access statistics, including quantity and geographical distribution of users, infrastructure usage level and user's appreciation, several Virtual Access metrics[28] have been defined and reported to WP13 for each reporting period.

Definitions:

- User: scientist/researcher running analytics tasks on ECAS (CMCC or DKRZ instance)
- Job: an analytics task run by a user on ECAS
- Baseline: the level achieved before the start of the reporting period (until M17). It is noteworthy that the values measured from project months M1 to M17 are also included into this count, besides the initial values before the project start.

The following metrics have been considered:

- Number of new registered users: number of new users registered to ECASLab, based on the ECAS internal accounting system.
- Number of active users: number of users who have run at least one analytics task over each reporting period, based on the ECAS internal accounting system.
- Number of jobs: number of jobs run on the ECASLab cluster, based on the ECAS internal accounting system.
- Number of cores hours: number of cores hours used on the ECASLab cluster, based on the ECAS internal accounting system.
- Number and names of the countries reached: number and names of countries reached, based on information provided by users for registration.
- Satisfaction: User satisfaction (6-point rating scale) based on feedback gathered by means of evaluation forms during the training events.

[26] https://www.medcordex.eu/cmip5_data_reference_syntax.pdf

[27] https://docs.google.com/document/d/1h0r8RZr_f3-8egBMMh7aqLwy3snpD6_MrDz1q8n5XUk/edit

[28] https://github.com/ECAS-Lab/ecas-accounting

- Marketplace views: number of views in the EOSC Marketplace, measured by Google Analytics (from Marketplace).
- Marketplace Orders: number of orders for the ECAS resource from the EOSC Marketplace, measured by Google Analytics (from Marketplace).

Metric	**Baseline Until May 2019 (includes Period 1 and 2)**	**Period 3 M18 - M26**	**Period 4 M27 - M39**
New registered users	147 (88 until end P1, 59 in P2)	36	51
Active users	134 (61 until end P1, 73 in P2)	74	87
Number of jobs	606K	1149K	~500K
Number of cores hours	6484	3170	2050
Countries	14	19	16
Satisfaction	4,6	5	5,2
Views	NA	32	99
Orders	NA	1	2

4.5 Lesson learnt

Following the roadmap defined at the beginning of the project, most planned integrations have been achieved and the ECAS thematic service has been successfully integrated into the EOSC marketplace and the EGI cloud infrastructure.

Beyond that, there are some key aspects that are worth considering for future development, integration and planning activities, especially regarding topics centred around user communities (user engagement, experience and interaction). The main considerations can be listed as follows:

1. The availability of a larger set of data and associated easy to use data catalog is a crucial point in order to deal with a more comprehensive set of needs and requirements for end-users; thus the new ECAS instance at DKRZ now directly integrates a 5 PByte data pool with important climate data collections for the user community (e.g. from CMIP5, CMIP6, Cordex) which are easily accessible via an intake[29] data catalog.
2. There has been an important technology shift in the last two years that largely impacted the climate science community: the availability of a community software stack promoted as part of pangeo[30], which supports conventional compute cluster-based deployments as well as cloud-based deployments (there are deployments accessible for the research community on google cloud as well as amazon). Additionally, this software stack addresses the wider

[29] intake data catalog: https://intake-esm.readthedocs.io/en/latest/

[30] Pangeo -- A community platform for Big Data geoscience https://pangeo.io/

geoscience community enabling cross community collaboration, so e.g. CMCC and DKRZ are engaged in EOSC-pillar with partners like IFREMER addressing both observational data and climate simulation data-based use cases.

3. Besides the core components, integration aspects with broad software ecosystems (i.e. Python toolboxes) would lead to a stronger support programmability and software-related users' needs; thus e.g. the pangeo developments mentioned before strongly influence the availability of xarray[31] based community software packages (supporting e.g. dask based parallelization).
4. The training program plays an important role in engaging with more users and expanding beyond the existing user base.
5. Bringing the community around to software development by using public repositories (i.e. GitHub) should be stressed more in the future. This would need higher-level tools for software publication and also results sharing.
6. A better integration with the marketplace could be useful in the future to simplify and automate the order placement, thus having a stronger impact on the exploitation of the results. For example, the users could be automatically redirected to the ECAS registration/login page after the order is placed/accepted.

4.6 Future plans beyond EOSC-hub

The ECAS thematic service will continue to operate under the EGI-ACE project in order to deliver compute and analytics capabilities to the end users. Specifically, compute capacity will be allocated on demand by the EGI-ACE IM/EC3 tool within the context of the EGI-ACE data spaces.

In addition, ECAS is also exploited in the H2020 EOSC-PILLAR project with a major focus on provenance management over data analytics experiments, thus enriching the current ecosystem with lineage information support.

ECAS is part of the IS-ENES compute service portfolio and, as part of this, it will be integrated with new compute backends at individual sites. Thus e.g. at DKRZ currently a parallel ECAS instance is available tightly integrating with the HPC environment and supporting xarray and dask based workloads, for which a growing user demand can be observed.

Data sharing based on “analysis ready data” on the cloud is growing. Thus DKRZ together with IS-ENES partners is currently working on data integration of ECAS with institutional (e.g. openstack based) as well as public cloud storage (e.g. amazon, google).

The availability of climate data pools near processing services (e.g. based on the OGC WPS standard) can be foreseen for the near future. IS-ENES partners including DKRZ and CMCC are working on WPS service offerings and DKRZ and IS-ENES partners will offer WPS compute services for data integration into the Copernicus climate data store. Thus WPS service integration into the ECAS

[31] xarray: N-D labelled arrays and datasets in Python http://xarray.pydata.org/en/stable/

thematic service and WPS service offerings based on the ECAS thematic service are important aspects partners are working on.

Besides that, the ECAS team will continue to support the thematic service carrying out new extensions and integration activities about the integrated components (e.g. JupyterHub interface, the Ophidia Framework and the related Python bindings).

Additional training courses (also as virtual and online events) are planned for the future with the aim of addressing new use cases coming from other research communities. Moreover, inter-thematic-service collaboration will be addressed in the future to discuss new integrated scenarios relying on multiple thematic services. In particular, based on preliminary interactions with OPENCoastS and the EGI team, some training events could be planned/organised as well as joint ECAS & OPENCoastS use cases could be evaluated in order to define possible common approaches.

5 GEOSS

5.1 Service Description

Service/Tool name	**GEO Discovery and Access Broker (GEO DAB)** **Virtual Laboratory (VLab)**
Service/Tool url	GEO DAB API: http://api.geodab.eu/ VLab API: http://vlabapi.geodab.org/
Service/Tool information page	GEO DAB: https://www.geodab.net/ VLab: https://essilab.wixsite.com/vlab
Description	GEO DAB is a key component of the GEOSS Platform, transparently connecting GEOSS User's requests to the resources shared by the GEOSS Providers. VLab allows the use of GEOSS datasets for the generation of new products, implementing all orchestration functionalities to ingest GEOSS data, execute workflows/models and save outputs.
Value proposition	GEO DAB goal is to simplify cross and multi-disciplinary discovery, access, and use (or reuse) of disparate data and information. GEO DAB is a brokering framework that interconnects hundreds of heterogeneous and autonomous supply systems by providing mediation, harmonization and transformation functionalities. VLab addresses the needs of scientists and modelers facilitating the generation of knowledge for an evidence-based decision-making process. It enables the execution of scientific models/workflows not only on EOSC but also on other cloud platforms, including AWS and the Copernicus DIAS.
Customer of the service/tool	Research Communities
User of the service/tool	N/A
User Documentation	GEO DAB: https://www.geodab.net/ VLab: https://confluence.geodab.eu/display/VTD/VLab+Documentation
Technical Documentation	GEO DAB: https://www.geodab.net/ VLab: https://confluence.geodab.eu/display/VTD/VLab+Documentation
Product team	CNR
License	N/A
Source code	N/A
Testing	Both GEO DAB and VLab utilize unit and integration testing. Testing framework is based on JUnit.

5.2 Initial ambition (in 2018)

The GEOSS (Global Earth Observation System of Systems) services support the implementation of the Sustainable Development Goals (SDGs) defined by the United Nations. Services scope is to help SDG monitoring and assessing by providing the necessary Indicators and Essential Variables (EVs) defined by the Community.

The initial ambition was to leverage EOSC capabilities (in particular IaaS and PaaS) to ensure high availability, reliability and scalability of the GEO DAB (Discovery and Access Broker) – one the core components of the GEOSS Platform.

5.3 Final software architecture and integration

The GEO Discovery and Access Broker (GEO DAB) is a key component of the GEOSS (Global earth Observation System of Systems) Platform, transparently connecting GEOSS User's requests to the resources shared by the GEOSS Providers. GEO DAB and its use of EOSC was described in more detail in D7.2 [R2]. In connection with Virtual Earth Laboratory (VLab), it is possible to use GEOSS datasets for the generation of new products: users can discover and execute workflows to generate new products useful for her/his analysis, utilizing the GEO DAB to discover and ingest input data and VLab to orchestrate and run the workflow of interest. VLab utilizes a Kubernetes cluster which was installed on EOSC infrastructure and utilized for the deployment of the VLab software modules. In addition to Kubernetes itself, VLab utilizes a set of ancillary cloud services (e.g. Web storage, queue service, etc.) which are provided as Kubernetes applications (e.g. MinIO, KubeMQ, etc.).

The EOSC provider of the utilized cloud services was CESNET-MCC (Czech Republic).

When the generation of a new product is requested, VLab implements the following steps: (i) retrieves the necessary input data utilizing GEO DAB APIs, (ii) stores the data in the execution environment, (iii) creates a Docker container with the required libraries and model source code for the product generation, (iv) launches the Docker container, and (v) saves the generated product making it available on the Web. Besides this default workflow, when possible, VLab can launch the Docker container for execution in a different Cloud environment which already hosts required input data; this allows a more efficient execution since it allows to avoid data download time. This was experimented with three DIAS platforms: ONDA, Sobloo and CREODIAS.

5.4 Impact and exploitation

The GEO DAB offers discovery and access to data provided to GEOSS by about 190 systems, with more than 400M discoverable metadata records. The usage statistics of the GEOSS Platform are maintained by GEOSS Platform team and will be published on the GEOSS Web Portal.

VLab provides the possibility to execute more than 20 scientific models, developed in different programming languages (including: Java, Python, NetLogo, Matlab, R) and over 1400 model executions.

A VLab-based demo was developed in the context of the EuroGEOSS Sprint-to-Ministerial activity. The demo enhanced the pilot presented in 2018 for the EOSC Launch event. A video was produced for the demo.

During the last GEO Plenary meeting in Camberra (November 2019), the developed demo was widely shown at the EC booth during the entire meeting, at the EuroGEO side event and during the plenary session. The general feedbacks were positive, particularly highlighting the value of being able to exploit different European cloud capacities (EOSC and DIAS platforms). It was however noted that a simplified GUI for decision-makers could improve the usability of the service. Besides, VLab was also utilized for the demo shown at the EOSC-hub week Closing Plenary.

Finally, a collaboration with EC JRC in the context of the Destination Earth activity developed a demo to show the enhanced multi-cloud support in VLab model executions.

5.5 Lesson learnt

The main gaps had been initially identified in D7.2 and D7.3. During the development of the project it was possible to consolidate the integration. This mainly concerned the use of non-cloud provided services for the web storage and queue messaging functionalities, e.g. deploying Kubernetes-based software. Although not optimal, this solution helped enabling a fully integrated Kubernetes-based deployment of VLab.

An important requirement for an efficient execution of models generating new products, especially in this Big Data era, is the possibility to use cloud-hosted data as inputs. In particular, the possibility to access the entire set of Copernicus Sentinel data natively in EOSC would be of great value.

CNR-IIA team (leader of the GEOSS TS) applied as Early Adopter for the EGI-ACE project where the use of EOSC for the described services will be continued and enhanced.

6 OPENCoastS

6.1 Service Description

Service/Tool name	**OPENCoastS**
Service/Tool url	https://opencoasts.ncg.ingrid.pt/
Service/Tool information page	http://opencoasts.lnec.pt/index_en.php
Description	The OPENCoastS service assembles on-demand circulation forecast systems for selected coastal areas and keeps them running operationally for a period defined by the user. This service generates daily forecasts of water levels and vertically averaged velocities over the region of interest for 48 hours, based on numerical simulations of the relevant physical processes.
Value proposition	A user-friendly web portal allows users to set up forecast systems of their coastal region of interest.
Customer of the service/tool	Coastal Research Communities, Coastal end-users, Coastal authorities
User of the service/tool	researchers, people in charge of emergency and daily planning at coastal infrastructures
User Documentation	Manual: http://opencoasts.lnec.pt/OPENCoastS_manual.htm Reference paper: Oliveira, A., A.B. Fortunato, J. Rogeiro, J. Teixeira, A. Azevedo, L. Lavaud, X. Bertin, J. Gomes, M. David, J. Pina, M. Rodrigues, P. Lopes, 2020. OPENCoastS: An open-access service for the automatic generation of coastal forecast systems, Environmental Modelling and Software, 124: 104585. DOI: 10.1016/j.envsoft.2019.104585
Technical Documentation	to be released by March 2021
Product team	LNEC, LIP, CNRS/LIENS, UNICAN
License	Apache License Version 2.0
Source code	to be released by March 2021
Testing	a suite of testing deployments, covering all options, was defined and is conducted every time a new release is performed

The OPENCoastS service builds on-demand circulation forecast systems for user-selected coastal sections anywhere in the world and maintains them running operationally for the time frame defined by the user. This service has three options, depending on the relevant physics:

1. 2D barotropic simulations - these simulations output water levels and depth-averaged velocities. Forcings are tides, wind, atmospheric pressure and river flow.

2. 2D barotropic simulations with wave-current interaction (2D W&C) - these simulations provide additional wave parameters. Wave-current interactions are simulated and forcings also include short waves. It can only be used in the North Atlantic.
3. 3D baroclinic simulations - these simulations provide 3D fields of velocity, salinity and water temperature, besides water levels. They can be forced by tides, river flow, temperature and salinity at all the boundaries, and also by the atmospheric surface forcing (wind, air temperature, pressure, humidity, solar radiation and down-welling longwave radiation). These forecasts can be generated anywhere in the world.

This freely available web service builds forecasts systems and maintains them in operation using the European Open Science Cloud infrastructure. Forecasts are based on the SCHISM community model's simulations using a computational grid uploaded by the user, to predict the 2D barotropic (with or without waves) or 3D baroclinic circulation in the coastal area of interest. The 2D waves and currents circulation option was added in the last few months.

The web platform includes the whole forecast workflow: system configuration, system management and forecasts viewer. Forecast accuracy can be controlled by the user at configuration stage by: i) the selection of the forcings from the several sources available (including Meteo France and NOAA's atmospheric predictions and ocean boundary conditions from CMEMS and FES2014, among others); and ii) the choice of several model parameters. The quality of the predictions can be assessed by automatic comparison with user selected EMODNET data stations. Default settings are also provided to facilitate the uptake of the OPENCoastS platform by people less familiar with the use of the SCHISM model. An open grid repository in Zenodo/GitHub is also available for testing the platform and to promote shared work.

OPENCoastS was also disseminated in peer-reviewed journals and books, including several presentations at international conferences. Two onsite training events were also promoted targeting coastal managers, Africa and East Europe: 14th Silusba and the 14th MEDCOAST conferences. A major final online training event was held in January 2021, where all service options were presented along with live demos. This training event also included a dedicated day for developers, promoting collaborations for future developments and setting the stage for opening the software to everyone.

More information can be retrieved from EOSC-hub deliverable 7.1, 7.2 and 7.3 [R1, R2, R3], on the several papers published in 2019: OPENCoastS core paper (published in open access), OPENCoastS grid repository paper and OPENCoastS 3D version papers [R4-R6]. A paper on the 2D wave and currents version is being prepared.

6.2 Initial ambition (in 2018)

OPENCoastS' initial ambition was to provide a service for setting up on-demand 2D circulation forecasts forced by waves and currents, applicable in the North Atlantic, fully integrated with core EOSC services. This service was expected to be delivered through a user-friendly web interface that simultaneously facilitates the task for non-experts, while making the use of e-infrastructures transparent to the user.

6.3 Final software architecture and integration

The OPENCoastS service architecture includes a frontend with a user interaction component for forecast systems configuration and management, via a web application, and a backend where models and mapping services run and a storage tier for preservation. The following figure illustrates this architecture with more detail.

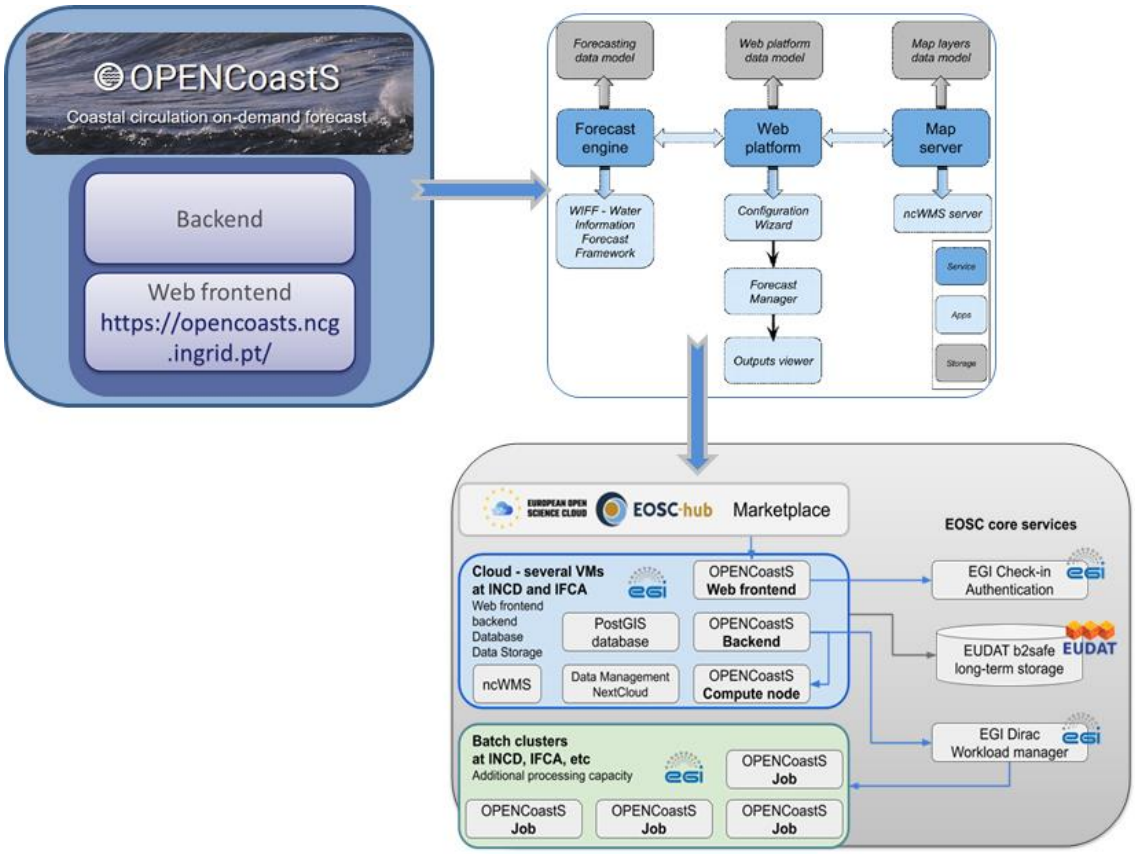


Figure 6.1: OPENCoastS service architecture.

Web application

This component provides access to the service, through web pages hosting wizard, manager and viewer applications (apps). Each of these apps allows the users to interact with the different aspects of the service while keeping them independent of each other. The Django Web Framework is used as the development basis of this component, which follows Django's design philosophy of having a project composed of applications, each with a set of concerns and functionalities.

Forecasting

While the web application, as the user facing component, allows the users to interact with the service and manage their operational real-time forecast systems, it is the forecasting component that is responsible for producing the forecast results. As a central piece, it interacts with all other components, directly or indirectly, to be able to gather all the necessary information to run simulations and make their results available.

Mapping

The mapping services complement the web application ones, by providing WMS services, which are then consumed by the viewer application. The ncWMS2 server is used to publish the spatial forecasting results on the web, sourced from netCDF files.

Storage

The storage component keeps the state of all services and is a requirement for all of them. The storage technologies range from typical relational databases servers to lower-level shared file systems. The relational database software used is the PostgreSQL, with PostGIS support, storing most structured information about the service. The object storage service, reviewed to be connected to EUDAT B2SHARE and B2DROP, makes available files used in and resulting from the simulations. Shared file systems, in this case NFS, are used to share folders and files among the computational jobs submission and mapping host resources.

Computation

The computational demands for the Opencoasts schisms require an infrastructure able to answer the HPC requirements for the forecasts.

EGI High Throughput Compute Thematic Service provides the distributed network of computing centres that allows to analyse the datasets and execute the parallel computing tasks leveraging the forecasts before the next day. The service delivers the workload, data management, integrated monitoring and accounting tools, allowing to manage all computational tasks and reporting the information about the availability and resource consumption.

Subsequently, the EGI Workload Manager Thematic Service provides the solution to manage and distribute the computing tasks maximising the usage of the computational resources. The service is based on DIRAC technology and is the main interface for submitting the jobs with a RESTful API.

Using this implementation Opencoasts can use more computational cores, not only on the Cloud infrastructure, but also from the computing GRID network.

The StoRM WebDAV service is used for data transfer between the Cloud and the computation infrastructure. It allows the access from the Opencoasts deployed services in the Cloud to the external computing environment, as defined in figure 2. With the RESTful interface provided by WebDAV implementation was possible to access the storage with common http requests, hiding the complexity of the storage services and the heterogeneity between different providers.

The udocker tool is used to run the processing scripts in the computing nodes with the required software dependencies and environment. The software is packaged inside a docker image to be launched with udocker in user space without requiring root privileges. The tool can be downloaded and executed entirely by the end user without requiring any type of privileges nor deployment of services by system administrators.

After setting into production all required services, EGI Accounting Thematic Service stores the user accounting records from the various services offered by EGI, in particular the Cloud resources, computing and storage usage. This information can be consulted online through the EGI Accounting Portal.

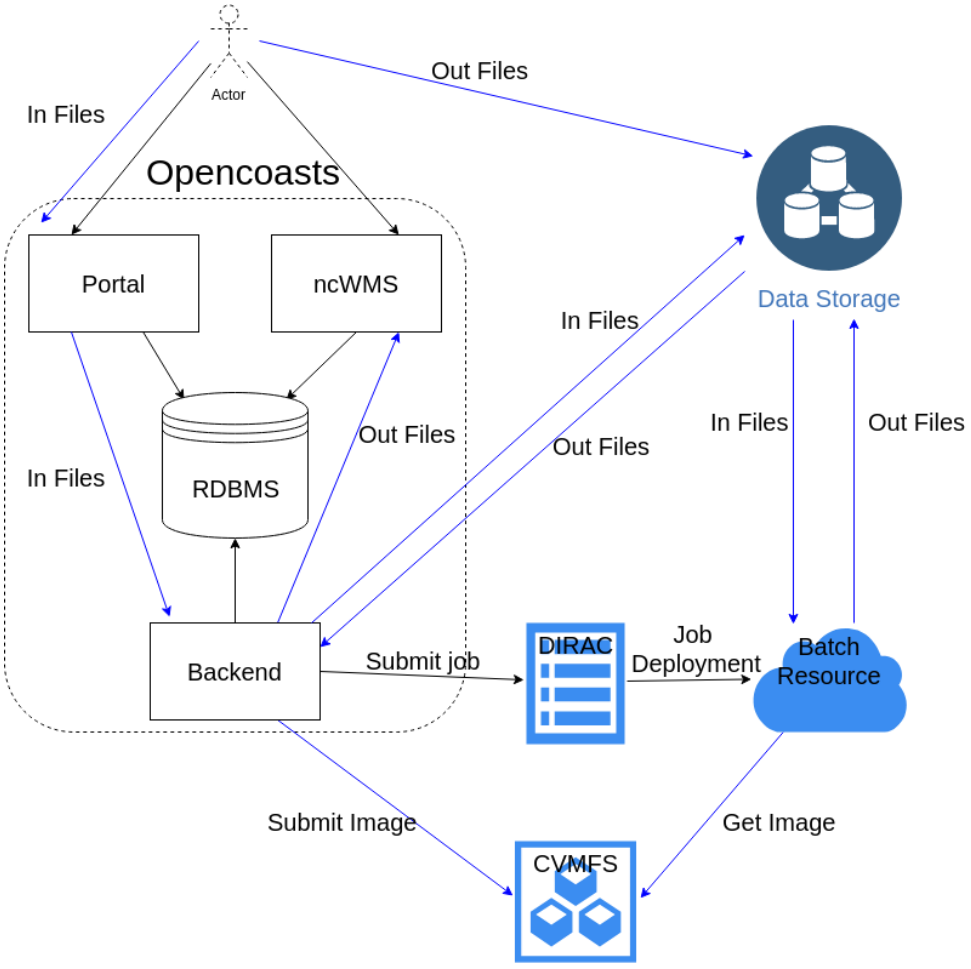


Figure 6.2: OPENCoastS integration for computational resources

6.4 Impact and exploitation

In the EOSC-hub project, OPENCoastS has grown from a national service to a worldwide platform with users on all 5 continents. Thanks to the core services provided by EOSC-hub, we have over 300 international deployments (applications of the OPENCoastS service to specific coastal sites) and users from 61 countries. Five international training events have spread the word on the advantage of using e-infrastructures and their core services to support user-oriented engineering platforms to address coastal research and management needs. The evolution of the service usage is summarized in the Table below.

	Baseline 2017	**Period 1** M6-M8	**Period 2** M9-M17	**Period 3** M18 - M27	**Period 4** M28 - End	**Total in project**
New users	0	18	165	51	158	392
New countries of users' origin	1	3	19	12	27	61
New deployments	0	38	97	52	150	337

After the end of the project, the following actions are planned:

- Extension to water quality predictions – the need to comply with several European Directives (such as the Water Framework Directive or the Bathing Water Directive) has prompted OPENCoastS users to request its extension to water quality variables. This extension is considered in the EGI-ACE proposal.
- Extension to hindcast runs – while forecasts have a large community of users and OPENCoastS has already been used for establishing a model in the EAP Taiwan service, full operationalization of hindcast runs is very important for the whole coastal community and can benefit greatly from the core services and infrastructure resources in the EOSC. This extension is considered in the EGI-ACE proposal.
- Integration with other EOSC services to provide added-value applications. An example is ECAS, where the requirements analysis to integrate its post-processing tool in OPENCoastS workflow was initiated in EOSC-hub.
- Allow to receive a broader users' community from OpenID Connect provided from EGI Check-in proxy service.
- Deploy the scripts environment in any EGI High-Throughput compute cluster using udocker to execute a container in user space without requiring root privileges.
- Promote dedicated training to users that require local installations of the service for high sensitive locations
- Promote dedicated training to end-users that have difficulty to generate the only input required by OPENCoastS: the computational horizontal and vertical grids.

6.5 Lesson learnt

The availability of core services and resources in EOSC has proven to be essential to catapult e-engineering services of importance and potential usage at world scale to their users. The experience of OPENCoastS and the very good feedback from its users have shown that the promotion of thematic service in e-infrastructures calls is the right path to promote these infrastructures in a very practical way that reaches the users by providing the needed services.

On a down-side, the complexity of integrating new services requires a deeper work between thematic services providers and core service personnel. This enhancement can be achieved with training for thematic services providers using tools and technical language that can be followed and understood by non-computer science people.

7 WeNMR

7.1 Service Description

Service/Tool name	**WeNMR**
Service/Tool url	https://www.eosc-hub.eu/services/WeNMR%20suite%20for%20Structural%20Biology
Service/Tool information page	http://www.wenmr.eu
Description	The WeNMR thematic services are a suite of web portals providing user-friendly access to complex computational workflows and tasks in the structural biology field.
Value proposition	The WeNMR portals allow inexperienced and experienced structural biologists to make use of state-of-the-art software for their research while benefiting from the computational infrastructure provided through the EOSC-hub project, in a transparent way.
Customer of the service/tool	Structural biology and bioinformatics communities, INSTRUCT-ERIC
User of the service/tool	Main users of the WeNMR are structural biologists of any degree of experience, with the aim of modelling and refining NMR structures, complexes of proteins and other biomolecules or fitting cryo-electron microscopy maps.
User Documentation	http://www.wenmr.eu
Technical Documentation	http://www.wenmr.eu
Product team	Utrecht University (UU), Consorzio Interuniversitario Risonanze Magnetiche di Metallo Proteine (CIRMMP), INFN Padova
License	All portals are freely accessible to non-profit users upon registration
Source code	N.A.
Testing	The portals are heavily used and have support mechanisms in place (e.g. via ask.bioexcel.eu). Updates to the portal machinery itself is subjected (in most cases) to continuous integration via GitHub/Jenkins.

The WeNMR thematic services are **a suite of web portals** providing user-friendly access to complex computational workflows and tasks in the **structural biology field**. The goal of these portals is to allow inexperienced and experienced structural biologists to make use of state-of-the-art software for their research while benefiting from the computational infrastructure provided through the EOSC-hub project, in a transparent way. As already described in Deliverables 7.1 and 7.2 [R1, R2], the WeNMR services make use of high-throughput computing (HTC) resources and some even of accelerated computing (GPGPUs) grid resources and cloud computing. Seven services vertebrate the WeNMR suite: AMPS-NMR (Nuclear Magnetic Resonance structure refinement), CS-ROSETTA (3D structure modelling of proteins), DISVIS (visualizing and quantifying accessible interaction space

in macromolecular complexes), FANTEN (computing the magnetic susceptibility anisotropy tensor for paramagnetic metalloproteins), HADDOCK (modelling complexes of proteins and other biomolecules), POWERFIT (rigid body fitting of atomic structures into cryo-EM density maps) and SPOTON (identification and classification of interfacial residues as hot-spots in protein-protein complexes).

7.2 Initial ambition (in 2018)

Since the very beginning of the project, the overall aim of the WeNMR services has been the provision of state-of-the-art, user-friendly services for the life science community, targeting researchers in the field of structural biology and structural bioinformatics. The strategy to achieve this was not limited to the provision of access to e-infrastructure but included implementing and developing state-of-the-art methods and software.

To tackle the above concept, the specific actions planned for WeNMR services within EOSC-Hub were:

1. Unification of the job submission mechanism for at least the most used portals in WeNMR:
2. Adoption of common SSO solutions for all the portals where a control of authentication mechanisms was desirable, going beyond specific solutions implemented in the structural biology domain;
3. Implementation of shared storage solutions for WeNMR users to manage their input data to the WeNMR portals and possibly also retrieve the output from calculations.

7.3 Final software architecture and integration

Most of the WeNMR portals have been in operation since several years. They are geographically located at the University of Utrecht in the Netherlands and at the University of Florence in Italy. All portals are web-based, built on a variety of technological solutions (e.g. Python, Flask, Apache, …), but all present a unified and well-recognizable front end to users. They make use of the EGI HTC resources to distribute jobs to the sites supporting the enmr.eu VO using in large majority of cases DIRAC4EGI for job submission. The GPGPU-grid enabled DISVIS and POWERFIT portals uniquely leverage udocker, a basic user tool to execute simple Docker containers in user space without requiring root privileges, developed by the INDIGO-DataCloud project and supported by EOSC-Hub. The adoption of udocker was crucial to enable execution of jobs on GPGPU resources located in Florence.

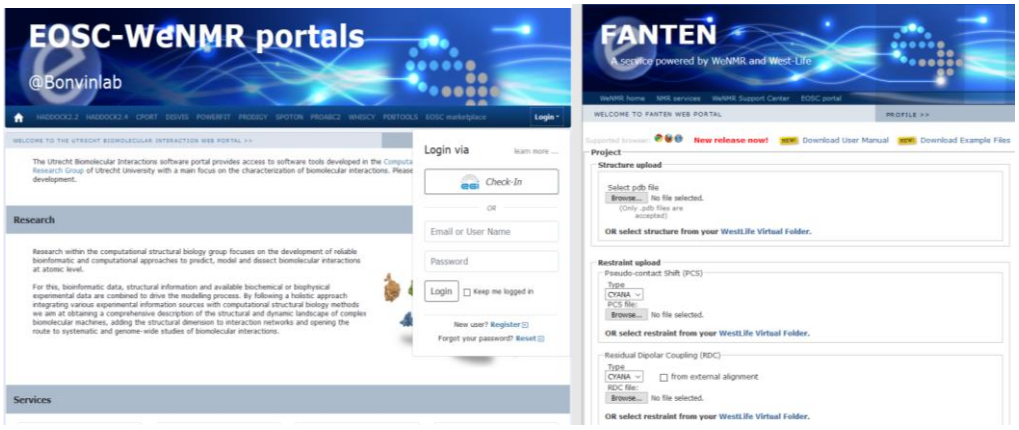


Figure 7.1: Front ends of the Utrecht WeNMR portals (left) and of the FANTEN web portal in Florence (right)

- **General architecture of the WeNMR portals**

All WeNMR portals are built on the same philosophy of shielding the end user from the complexity of accessing/using HTC (grid or cloud) resources. From a user perspective, a user only interacts with web-based portals, filling in forms and uploading data. Upon successful submission those data are processed through complex workflows calling typically a variety of software and using a combination of both local and EOSC HTC resources. Finally, the results are post-processed and presented to the user in a user-friendly manner, facilitating their interpretation. The general architecture is illustrated in the following figure.

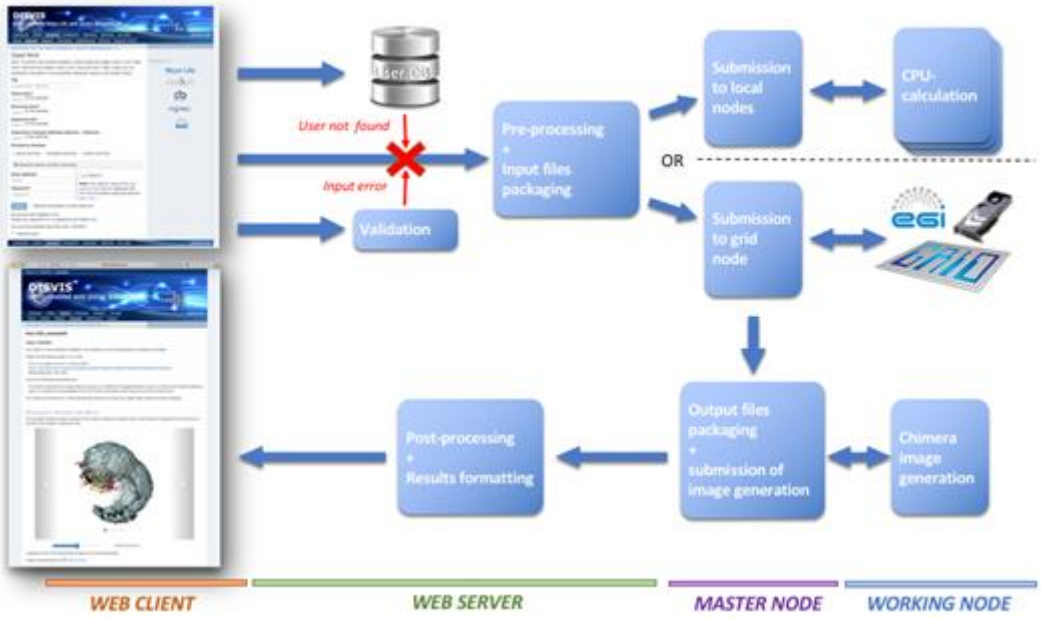


Figure 7.2: Illustration of the general workflow behind the WeNMR web portals

EOSC-Hub Integration

From day 1 of the project, the WeNMR Thematic Services have been in operation, sending over the first eight months of the project over 5 million jobs to the EOSC HTC resources, most of which through the DIRAC4EGI service. According to the EGI Accounting Portal, these account for over 15 million HS06 CPU Time hours consumed. Although the WeNMR services operate making use of opportunistic computing resource, the access to the federated resources of EGI has been formalized by a Service Level Agreement (SLA) between EGI.eu and the enmr.eu VO (represented by the Faculty of Science – Department of Chemistry of Utrecht University). This SLA was established in 2016 and has been renewed in 2018, granting to enmr.eu VO until 31/12/2020 an amount of opportunistic computing time up to 53 Million of normalized CPU hours and opportunistic storage capacity up to 54 TB[32]. Five resource centres signed this last version of the SLA: INFN-PADOVA (Italy), TW-NCHC (Taiwan), SURFSara and NIKHEF (The Netherlands), NCG-INGRID-PT (Portugal).

During the first nine month of the project, a number of WeNMR portals have been migrated from the old gLite-based job submission to the EOSC-hub **DIRAC4EGI** service. Further, all portals are now offering Single Sign On (SSO), either through the legacy West-Life SSO which connects to both ARIA (the access management solution of the Structural Biology infrastructure INSTRUCT-ERIC) and the old legacy WeNMR SSO[33], or directly through the **EGI Check-in**. Users can now register and use the WeNMR services using the Check-in, allowing them to use a variety of identity providers.

INFN has been hosting an instance of the EOSC-hub OneData service, offering up to 10 TB of storage space to the WeNMR community. WeNMR users can request a storage space at the Oneprovider service hosted at INFN-Padova data center by connecting to the Onezone server (onezone.cloud.cnaf.infn.it) hosted at INFN-CNAF data center (located in Bologna). INFN developed a new plugin to enable the West-Life SSO as authentication method for Onedata. The OneData space was integrated with the Virtual Folder (VF), a tool developed in the context of the West-Life project. Currently integrated in several WeNMR portals, it acts as a gateway for many storage systems, such as Dropbox, EOSC B2Drop and any other system accessible through the WebDAV protocol. INFN developed a plugin for integrating VF with Onedata, i.e. to enable Onedata storage system as an additional back end. However, there has not been a systematic use in production of this integration.

7.4 Impact and exploitation

The impact of the WeNMR services can clearly be seen in the number of countries reached which increased during the project from 96 at the start to more than 120, demonstrating worldwide impact.

[32] https://documents.egi.eu/public/ShowDocument?docid=2751

[33] https://indico.egi.eu/event/3903/sessions/2838/attachments/8596/9939/EOSC_HUB_Malaga18.pdf

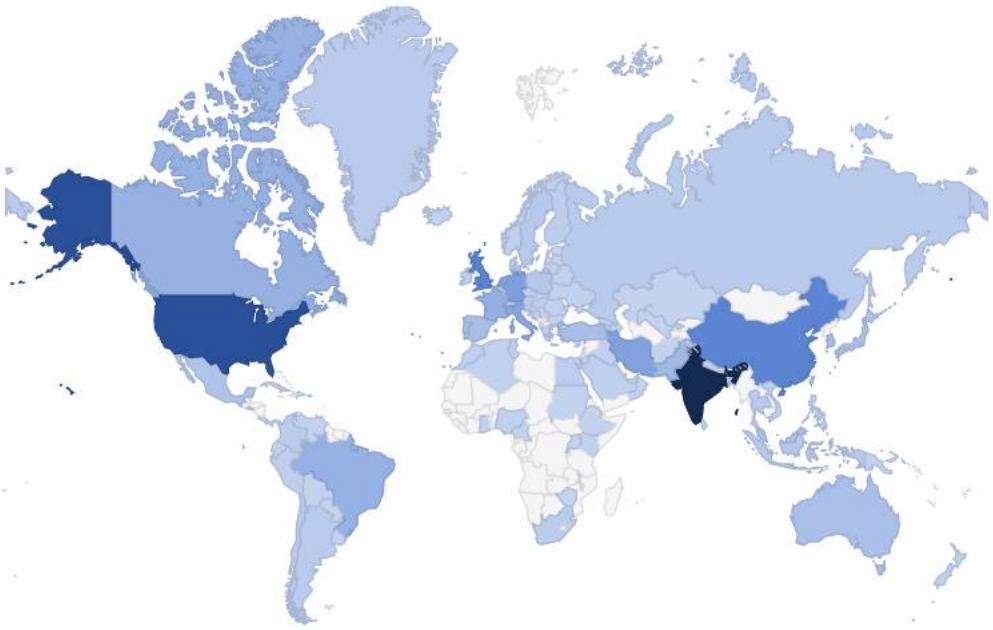


Figure 7.3: Worldwide distribution of WeNMR registered users.

The number of newly registered users per reporting period has also steadily increased during the project, from a defined based line of 1750 to more 4300 per nine months (reporting period) (Figure 7.4). The total number of WeNMR registered users has passed last December 20’000, an impressive number. All together the WeNMR portals have consumed a large amount of HTC resources over the last three years, as clearly shown in Figure 7.4. Note in particular the increased usage during the COVID19 first lockdown period of 2020. This increased CPU consumption is the result in particular of an increased use of the WeNMR HADDOCK portal, as reflected in the number of user submissions per month in the following figure. Also note the steady fraction of COVID19-related submissions that we have started monitoring since April 2020. Each user submission translates into hundreds to thousands of HTC jobs submitted to the EGI workload manager (DIRAC4EGI).

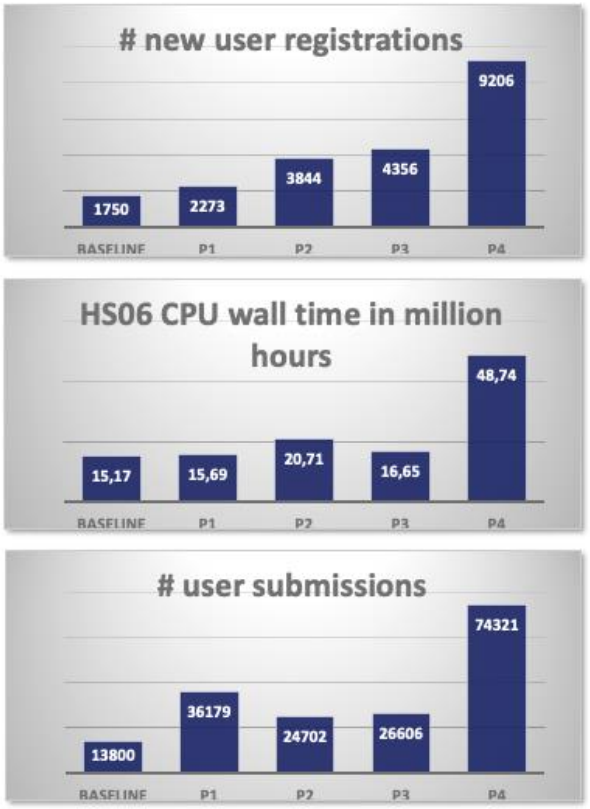


Figure 7.4: Main WeNMR KPIs over the different reporting periods.

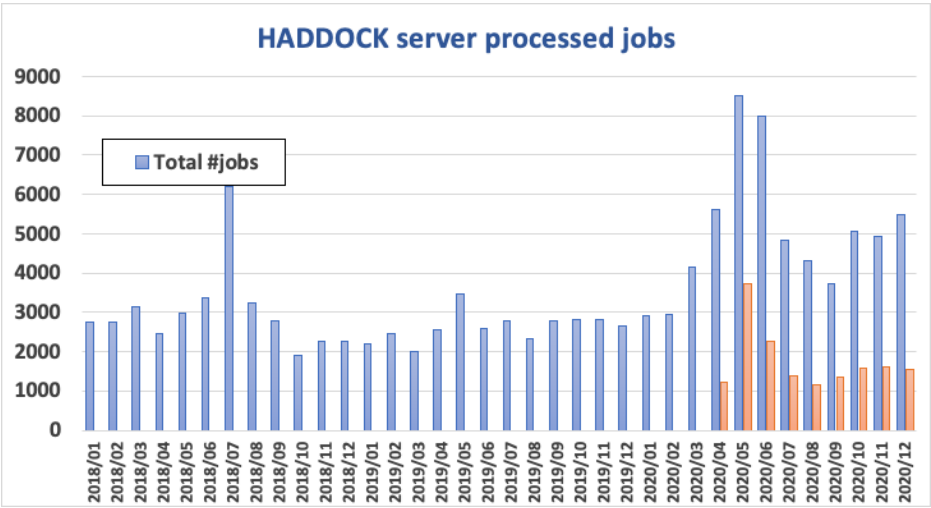


Figure 7.5: HADDOCK WeNMR portal user submission per month. Blue bars indicate the total number of submissions and red bars the COVID19-related submissions.

7.5 Lesson learnt

The following list recapitulates some aspects observed during the work related to WeNMR services in EOSC-Hub that are worthy of consideration also as a contribution to shaping future initiatives.

1. the day-to-day interaction with the end users (i.e. the scientists in their research labs) remains crucial to foster the adoption of innovative tools. In this respect it is also important to keep in mind that in a lively community there is always a significant turnover of users. This happens because end users may no longer need (some) services when they move on to new tasks, e.g. when a funded project ends, or jobs and consequently this creates the need for continuously engaging and training new users. An additional aspect is that even within mature communities, such as structural biology, technological and methodological development is constantly changing usage scenarios and requires corresponding innovation in the software tools provisioned. All these novelties must be timely communicated to the users and support from the developers is required in order to enable optimal exploitation of the services
2. The use of the marketplace may not be straightforward for services that are provided on a free basis to (academic and non-profit) users. The general expectation in the structural biology/bioinformatics community is that free services should be used preferentially if not exclusively. As a consequence, providing free services next to for-pay services in the same platform proved confusing and was not received favourably especially by new users. We should also note that the majority of users are finding the WeNMR portal directly and not via the marketplace.
3. Word-of-mouth and dissemination activities at community's events remain the best ways to engage scientists, next to publications in scientific journals.
4. Although cloud storage has become a routine commodity for everybody in our everyday lives, it was not obvious to implement a seamless way of using cloud storage for the WeNMR services that suited individual users. In fact, although the possibility to exchange data via cloud storage was indeed implemented in our portals it was not really exploited by the community and its usage remained limited to selected case studies, particularly when they involved large or particularly complex data sets. This is probably due to the fact that the majority of real-life applications in structural biology require moving only a handful of small files or involve very large files (e.g. in cryo-EM) for which current cloud storage solutions are not practical.
5. Although the EGI Check-In was implemented in the portal as a SSO mechanism, we have only seen very limited usage. The majority of users seem to prefer to register directly into the WeNMR portal by creating new credentials rather than using the SSO.

8 DARIAH

8.1 Service Description

Service/Tool name	**DARIAH Science Gateway**
Service/Tool url	https://dariah-sg.irb.hr
Service/Tool information page	https://dariah-sg.irb.hr
Description	The DARIAH Science Gateway is a web-oriented portal, developed during the EGI-Engage project as part of the DARIAH Competence Centre and is specially tailored for the researchers coming from the Digital Arts and Humanities disciplines. It currently offers several cloud-based services and applications that can be accessed via the portal. Among others, this includes Semantic and Parallel Semantic Search Engines (SSE and PSSE), DBO@Cloud, Workflow Development, and support for several file transfer protocols.
Value proposition	Provision of a centralized portal for various Digital Arts and Humanities services connected with the EGI AAI. Easy integration of new services and high-level abstraction for accessing and exploiting cloud computing infrastructure on IaaS level for inexperienced users.
Customer of the service/tool	Digital Arts and Humanities research groups and projects
User of the service/tool	Individual researchers, small to large research groups
User Documentation	DARIAH Science Gateway help pages: https://dariah-sg.irb.hr/help2 Simple Cloud access service: https://ltos-gateway.lpds.sztaki.hu/web/wizard/help Gutenberg use case for Simple Cloud access service: https://bitbucket.org/davordavidovic/textanalysis/src/master/
Technical Documentation	-
Product team	RBI, SZTAKI
License	The gateway is freely accessible to non-profit users upon registration
Source code	gUSE Grid and Cloud Science Gateway: https://sourceforge.net/projects/guse/
Testing	-

Service/Tool name	**Invenio-based Repository**
Service/Tool url	https://indigo-paas.cloud.ba.infn.it
Service/Tool information page	https://dariah-portal.cloud.ba.infn.it

Description	The Invenio-based Repository is a service that enables researchers and scholars to easily create, deploy, and configure their own Invenio-based repository and host it on the EGI Federated Cloud infrastructure. The service aims at smaller research groups lacking adequate technical support and the budget to acquire their own infrastructure for hosting data repositories.
Value proposition	Easy creation of new (empty) instances of the Invenio-based Repository in the cloud. The initialization and setup of a new repository instance do not require technical knowledge with respect to configuring neither Invenio nor cloud infrastructure.
Customer of the service/tool	Digital Arts and Humanities research groups and communities. General research communities that do not have access to their own repositories.
User of the service/tool	Research group, research projects in the domain of Digital Arts and Humanities.
User Documentation	Deploying open-data repository in the Cloud (video): https://www.indigo-datacloud.eu/deploy-open-data-repository-cloud-using-marathon
Technical Documentation	-
Product team	RBI, Indigo-DataCloud
License	GNU General Public Licence 2
Source code	DockerHub container: https://hub.docker.com/r/indigodatacloudapps/dariah-repository Source code: https://github.com/indigo-dc/dariah-repository Ansible roles: https://github.com/indigo-dc/ansible-role-dariahrepo
Testing	-

Service/Tool name	**DARIAH-DE Repository**
Service/Tool url	https://repository.de.dariah.eu/publikator/
Service/Tool information page	https://de.dariah.eu/en/repository
Description	The DARIAH-DE repository is a service and an easy-to-use storage system for research data from the Digital Arts and Humanities that can store, modify, annotate, search, and access a large amount of diverse data, both structured and unstructured. The repository is based on the Common Data Storage Architecture (CDSTAR), a generic research data repository and archival service. CDSTAR integrates the ability to store metadata along with the research data in a flexible metadata schema that can be tailored to the specific use in different scientific disciplines. Additionally, the data objects that are stored in CDSTAR can be registered automatically at the ePIC Persistent Identifier (PID) service. A role-based security concept is

	also integrated into CDSTAR, which allows the protection of data sets with an individual set of permissions and rights for each user. Additionally, CDSTAR is capable of using SAML infrastructures such as the one provided by DARIAH in order to verify data access. This allows to use of the data permissions for the infrastructure at a central data point provided by DARIAH and to use the DARIAH access credentials.
Value proposition	Integration of the DARIAH-DE Repository with the EGI AAI allows users to directly access the service. Integration of the DARIAH-DE Repository with other services from EOSC-related projects increases the user community and increases the range for data re-use.
Customer of the service/tool	Digital Arts and Humanities communities
User of the service/tool	Individual researchers, research groups, institutions, research projects
User Documentation	https://repository.de.dariah.eu/doc/services/submodules/publikator/docs/index.html
Technical Documentation	http://repository.de.dariah.eu/doc/services/
Product team	GWDG
License	Apache License, Version 2.0 GNU LESSER GENERAL PUBLIC LICENSE, Version 3
Source code	The DARIAH-DE Repository consists of three main modules. The source code can be found at the respective repositories: Publikator: https://projects.gwdg.de/projects/publikator/repository DH-crud: https://gitlab.gwdg.de/dariah-de/dariah-de-crud-services DH-Publish: https://gitlab.gwdg.de/dariah-de/kopal-library-of-retrieval-and-ingest
Testing	https://dhrepworkshop.de.dariah.eu/

The DARIAH Thematic Service provides user-friendly solutions, i.e. services and applications, primarily addressing the needs of different research groups within the DARIAH-ERIC community and the Digital Arts and Humanities research domain in general. From the multitude of requirements communicate by these communities, the DARIAH Thematic Service provides a set of web-based services enabling end-users to seamlessly store, describe, and share their datasets, to discover, browse, and reuse datasets shared by others, and to perform elemental analysis on those data. The DARIAH Thematic Service is freely accessible to members of the DARIAH-ERIC community. For more details on the respective services please refer to Deliverables D7.1 and D7.2 [R1, R2].

The DARIAH Thematic Service provides three independent research services:

1. the DARIAH Science Gateway,
2. an Invenio-based Repository, and

3. the DARIAH-DE repository.

The DARIAH Science Gateway is a web-oriented portal that offers several cloud-based services and applications specifically tailored to the needs of the Digital Arts and Humanities communities. The portfolio accessible through the DARIAH Science Gateway includes the Semantic Search Engine (SEE), the Parallel Semantic Search Engine (PSSE), the Cloud-based repository of Bavarian dialects (DBO@Cloud), and the Simple Cloud Access and Workflow development. Furthermore, the gateway supports several file transfers protocols.

The Invenio-based Repository is a PaaS-type repository-in-the-cloud based on Invenio. This is a service that supports data owners with no access to an institutional repository to easily create, deploy, and configure Invenio-based repository instances and host them on a cloud infrastructure.

The DARIAH-DE Repository is a repository service for human and cultural research data. The DARIAH-DE repository is one of the central components of the DARIAH-DE Research Data Federation Infrastructure, which aggregates various services and applications and can be used comfortably by DARIAH ERIC users.

8.2 Initial ambition (in 2018)

Since the beginning of the project, the initial goal of the DARIAH Thematic Service was to support the Digital Arts and Humanities research communities in DARIAH and beyond in terms of service provisioning, access to infrastructure, and support for research communities. In addition, a further goal was to bring not only EGI resources and services to those communities but also those provided by INDIGO-DataCloud and EUDAT as well. The work plan continues the activities started during the EGI-Engage DARIAH Competence Centre and focuses on increasing the number of DARIAH-EU research communities actively exploiting the pan-European research resources. The three services already described above, i.e. the DARIAH Science Gateway, the Invenio-based Repository, and the DARIAH-DE repository, were planned to be maintained and provided to the end-users during the EOSC-hub project.

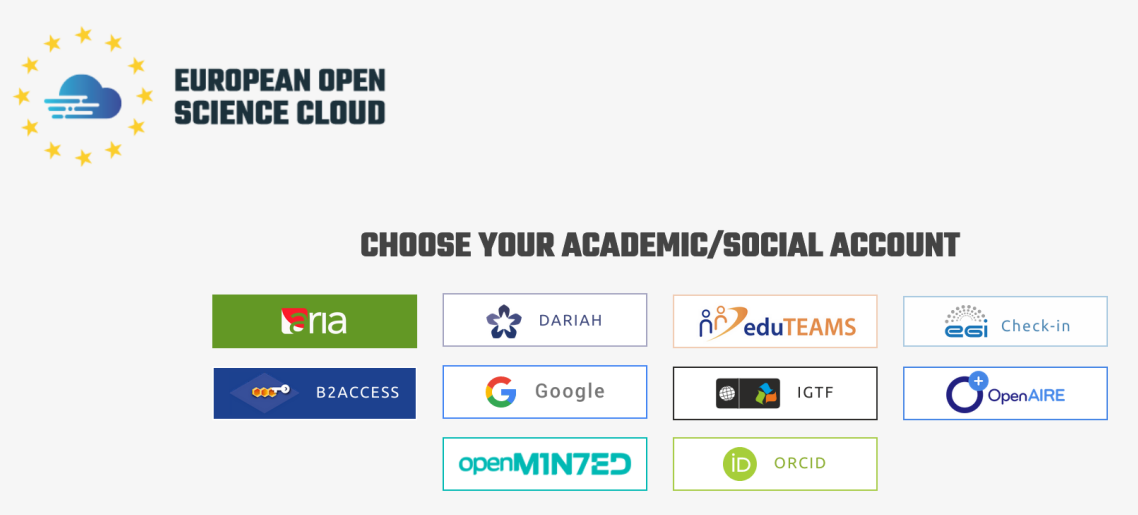
The following actions were envisaged at the beginning of the project:

- Maintain the DARIAH TS services,
- increase the number of individual users of the DARIAH TS services and the end-user, coming, from Digital Arts and Humanities research domains, of EGI/EUDAT/Indigo-DataCloud services and resources,
- ensure compute and storage resources to host the maintained services, and
- increase the DARIAH TS portfolio with new end-user services.

8.3 Final software architecture and integration

The software architecture for the three services has been described in detail in D7.2 “First report on Thematic Service architecture and software integration” [R2], which has been published in January 2019. From an architecture point of view, the document represents the final status, although the implementation of the services has been further evolved.

In general, all services maintained by the DARIAH Thematic Service are accessible through the DARIAH AAI, which is integrated with the EOSC AAI. This helps the target communities to identify with EOSC and to add further services relevant to them.



DARIAH Science Gateway

DARIAH Science Gateway is based on the WS-PGRADE/gUSE gateway technology and the Liferay portlet container framework. Together with the customization methodologies of WS-PGRADE/gUSE, the application/service developers can easily create user-friendly portlets. The gateway provides an API for creating portlets, which are using the services of WS-PGRADE/gUSE, called the Application-Specific Module (ASM API). This API enables application developers to call services of WS-PGRADE/gUSE for importing, modifying, and running existing workflows. Parameter-sweep job wizard enables users to run their application processing large input data step by step following six simple steps: executable upload, static input upload, parameter-sweep input upload, command-line argument definition, resource selection, and definition of output files. Finally, Data Avenue is a set of services enabling users to manage their data located on remote storage resources easily. The set of available operations includes browsing, directory creation/renaming/removal, file up-and

download, file management (rename/remove/move), and file transfer between different types of storage.

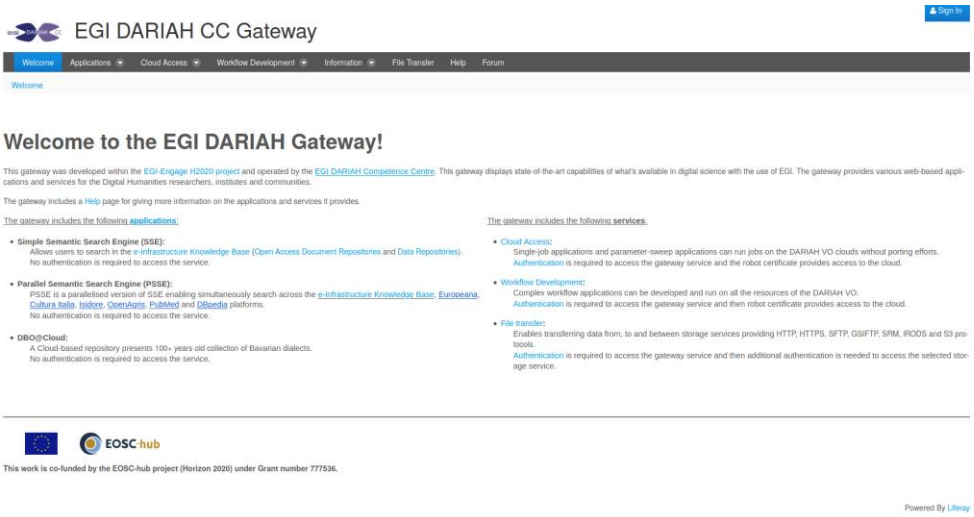


Figure 8.1: Welcome page DARIAH Science Gateway

The core technology of the DARIAH SG is based on the following components:

- WS-PGRADE/gUSE science gateway frameworks,
- Application-Specific Module (ASM),
- Parameter-sweep job wizard,
- DataAvenue.

The architecture of the DARIAH Science Gateway includes the following components:

- DARIAH CC VO
- DARIAH CC portal hosting:
    - eduGAIN login module
    - Application portlets
    - Workflow development portlets
    - ASM API
    - Cloud access portlet
    - File transfer portlet

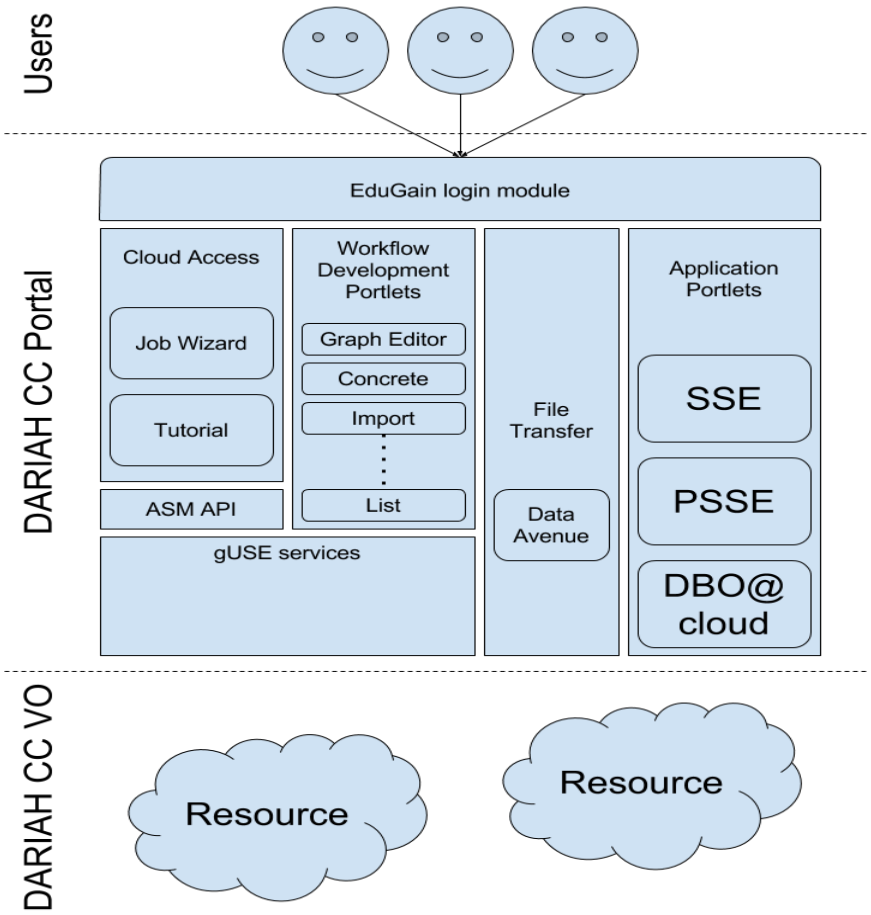


Figure 8.2: DARIAH Gateway architecture

Invenio-based Repository

The Invenio-based repository services main functionality is to provide an easy deployment of an invenio-based repository in the Cloud. The user interface is based on FutureGateway which provide functionality such as basic configuration of the repository, e.g. required resources in terms of the number of CPUs and storage size and fine-tuned (optionally) configuration of the Mesos cluster size and per-sub-service resource allocation. The user then starts the process of the repository creation in the cloud. The user request is submitted as a TOSCA template to the PaaS Orchestrator that coordinates resource provisioning and virtual infrastructure deployment. Once deployed virtual machines are automatically configured running dedicated Ansible roles shared on Ansible Galaxy. The Invenio components are installed via Ansible and executed as long-running services on top of a Mesos cluster through its framework, Marathon, to ensure fault tolerance and scalability. The IP address of the newly created repository instance is returned to the user via FutureGateway. Afterwards, the end-user can access the repository using the provided IP.

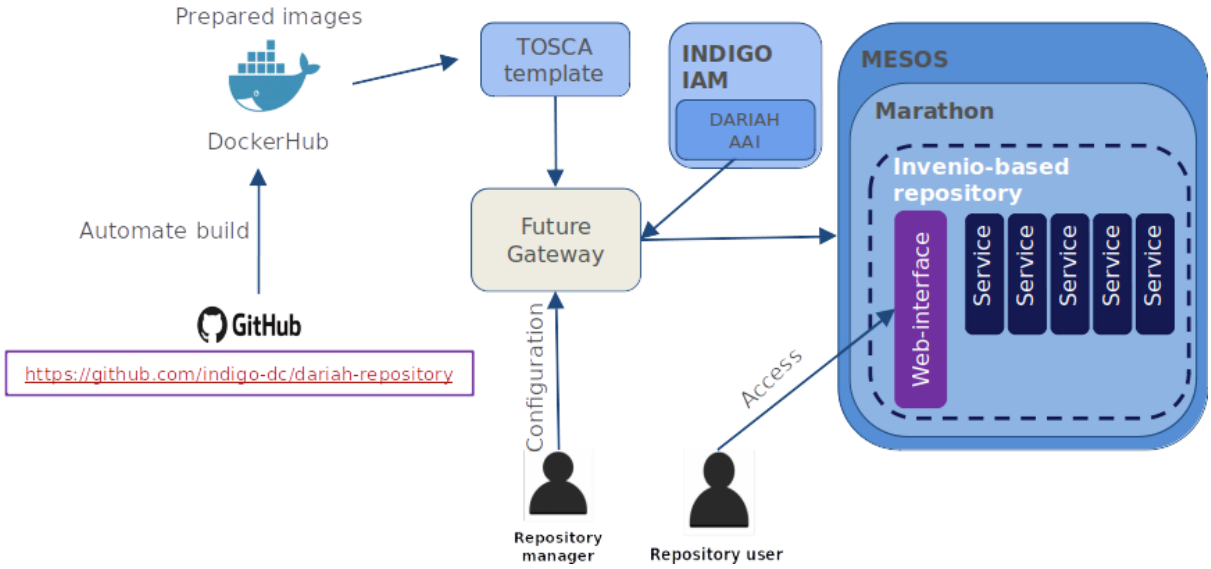


Figure 8.3: Schematic description of the Invenio-based Repository service

DARIAH-DE Repository

With respect to the DARIAH-DE Repository, three major strands of work have been followed: (i) the evolution and maintenance of the service, (ii) the integration with services and solutions from other EOSC-related projects, and (iii) the certification of the repository.

The service has been continuously improved, and the following releases have been published: 1.4.2 (Dec. 2017, baseline for the development), 2.0.8 (Mar. 2020), and 4.0.16 (Oct 2020). The later releases follow the architecture as referenced above. The documentation and the training material have been updated accordingly.

The DARIAH-DE Repository has been a central component of integration efforts with projects within the EOSC landscape. In collaboration with OpenAIRE, an interface has been specified and designed to enable the harvesting of the DARIAH-DE Repository through OpenAIRE. The full implementation has not yet been achieved but is planned to be implemented in 2021. Furthermore, the DARIAH Thematic Service collaborated with the SSHOC project and integrated the CLARIN Switchboard (see also Section 2). This integration enables users of the DARIAH-DE Repository to choose between a variety of TDM services and automatically apply them to a chosen data set. With just one click, the respective part of a collection is fed into the respective service, which makes reusability as easy as possible.

Last, but not least, the certification of the DARIAH-DE Repository according to the CoreTrustSeal (CTS) has been started. Statements regarding most of the CTS have been written and discussed with the certification team at Göttingen. A submission of the final certification request is planned for Q2/2021.

8.4 Impact and exploitation

In general, the DARIAH Thematic Service has been actively promoting its services and participated in a variety of outreach and exploitation activities. This includes common channels like Twitter, but also blogs and websites related to the Digital Arts and Humanities.

DARIAH Science Gateway and Invenio-based Repository service show a steady increase in the number of users from DARIAH community. However, both services can be used as general services for much wider research communities. This is demonstrated by the increased number of visitors from outside Europe, especially from the USA and Asia. In total, 1700 users from 76 countries worldwide have visited the services. Since the growing interest in the services, new applications (portlets) are planned to be included in the gateway as well as extending the portfolio of the DARIAH TS with new standalone services, with the focus on data processing and analysis tools, such as JupyterLab.

DARIAH-DE Repository

With respect to the DARIAH-DE Repository, the DiscussData project (https://discuss-data.net/) has to been particularly mentioned as its repository, which is offered to researchers focussing on the post-Soviet region, has been designed and implemented as a specific front-end to the DARIAH-DE Repository. This demonstrates the general applicability of the taken approach. The DiscussData repository can be easily used by a broad number of users thanks to the integration with the EOSC AAI.

The success of the outreach is also reflected in the numbers:

	Baseline 2017	**Period 1** M1-M8	**Period 2** M9-M17	**Period 3** M18 - M27	**Period 4** M28 - End
Number of collections	./.	0	70	117	165
Number of objects/DOIs	./.	0	566	836	1140

Overall, in particular in Germany, the interest in using the DARIAH-DE Repository has been growing, as it is already recommended by funding agencies as a target repository for projects from the Digital Arts and Humanities communities.

8.5 Lesson learnt

The Digital Arts & Humanities communities are in constant need of new digital tools and services. Thus, the DARIAH ERIC and the national counterparts support the communities through a large portfolio of services, training materials, data collections, and more. The collaboration with and contribution to EOSC-hub is seen as a great benefit from the DARIAH perspective and, among other things, led to large endeavours like e.g. the SSHOC project, which took up results and findings from the DARIAH Thematic Service (and its predecessor, the EGI-Engage DARIAH Competence Centre)

and evolve them further into a rich, integrated service portfolio for the Digital Arts & Humanities within EOSC. What has been, in general, an obstacle for the faster integration with EOSC are the different infrastructures and DevOps approaches taken by DARIAH compared to EOSC. This made the integration in some parts difficult. Furthermore, the effort of bringing services in the EOSC Market Place was partially underestimated by the DARIAH Thematic Service and should have been approached earlier in the project.

9 LifeWatch

9.1 Service Description

Service/Tool name	**Plant Classification**
Service/Tool url	http://deep.ifca.es/plants/
Service/Tool information page	https://github.com/IgnacioHeredia/plant_classification
Description	This service has been developed by IFCA-CSIC. It consists of a large-scale plant classification algorithm based on the ResNet convolutional neural network architecture.
Value proposition	This tool can definitely open the field to active contributions of non-expert users including citizen scientists.
Customer of the service/tool	http://deep.ifca.es/static/images/logoIFCA.png
User of the service/tool	It is a citizen collaboration tool. It is an open API that allows access to any user: those with a special interest in nature and botany, but also those curious people that want to know about a specific plant at some moment. Of course, its use is also aimed at developers and researchers
User Documentation	https://github.com/IgnacioHeredia/plant_classification
Technical Documentation	https://github.com/IgnacioHeredia/plant_classification
Product team	LifeWatch, IFCA-CSIC
License	https://github.com/IgnacioHeredia/plant_classification http://www.apache.org/licenses/LICENSE-2.0
Source code	https://github.com/IgnacioHeredia/plant_classification
Testing	-

Service/Tool name	**Glacier Lagoons of Sierra Nevada**
Service/Tool url	https://lagunasdesierranevada.es/
Service/Tool information page	https://lagunasdesierranevada.es/
Description	This service focuses on citizen collaboration and bio-conservation and constitutes the end result of a Citizen Science Campaign (“74 high mountain glacier oasis”) created and coordinated by the Department of Ecology of the University of Granada with the collaboration of the Sierra Nevada National Park, and the Global Change Observatory of Sierra Nevada. Its main aim is to involve society in the investigation and protection of the high mountain sites of Sierra Nevada.

Value proposition	The potential users of this service are mountaineers and citizens with a general interest in nature preservation. Counting on the voluntary participation, the researchers from the UGR seeks to enlarge the historical record of lagoon photographs and share all the scientific and practical information about the conservation, science and practices around these vulnerable ecosystems
Customer of the service/tool	http://secretariageneral.ugr.es/pages/ivc/descarga/_img/vertical/ugrmarca01color_2/!
User of the service/tool	It is a platform for citizen collaboration in which anyone can participate by sharing their photos. The photos are from an area located in Granada, Sierra Nevada, so the area of collaboration is limited. The collaborators (photographers) are mainly lovers of mountain and ecology.
User Documentation	-
Technical Documentation	-
Product team	LifeWatch, University of Granada
License	https://lagunasdesierranevada.es/condiciones-de-uso/
Source code	-
Testing	-

Service/Tool name	**GBIF Spain Spatial Portal**
Service/Tool url	https://espacial.gbif.es/
Service/Tool information page	https://espacial.gbif.es/
Description	GBIF.es data access under biogeographic context provides access with advanced facets to GBIF biodiversity data under a biogeographic context. Map service, to visualise and analyse relationships between species, location and environment.
Value proposition	Map service, to visualise and analyse relationships between species, location and environment.
Customer of the service/tool	https://www.gbif.es/wp-content/uploads/2017/05/gbif-logo.svg
User of the service/tool	This tool allows, through a map, to access the species of an area, load lists of species, generate graphs, make classifications and even create data sets.
User Documentation	https://www.gbif.es/en/portal-nacional-de-datos/
Technical Documentation	https://www.gbif.es/en/portal-nacional-de-datos/
Product team	GBIF Spain, LifeWatch
License	https://www.gbif.es/en/politica-de-privacidad/

Source code	-
Testing	-

Service/Tool name	**GBIF Spain Species**
Service/Tool url	https://especies.gbif.es
Service/Tool information page	https://especies.gbif.es
Description	GBIF.es data access under biogeographic context provides access with advanced facets to GBIF biodiversity data under a biogeographic context. Aggregates data on species description, common names, taxonomy, image gallery, sequence data and bibliography.
Value proposition	Aggregates data on species description, common names, taxonomy, image gallery, sequence data and bibliography.
Customer of the service/tool	https://www.gbif.es/wp-content/uploads/2017/05/gbif-logo.svg
User of the service/tool	It is a search tool where researchers can search for descriptions, common names, images, ... The content can be downloaded. It is also useful for students. The database contains species from all over the world, not only from Spain; therefore it is not only focused on research on Spanish species.
User Documentation	https://www.gbif.es/en/portal-nacional-de-datos/
Technical Documentation	https://www.gbif.es/en/portal-nacional-de-datos/
Product team	GBIF Spain, LifeWatch
License	https://www.gbif.es/en/politica-de-privacidad/
Source code	-
Testing	-

Service/Tool name	**GBIF Spain Species List**
Service/Tool url	https://listas.gbif.es/public/speciesLists
Service/Tool information page	https://listas.gbif.es/public/speciesLists
Description	GBIF.es data access under biogeographic context provides access with advanced facets to GBIF biodiversity data under a biogeographic context. This service permits to upload your own list of species to use in the GBIF.ES portal or use a list uploaded by other GBIF.ES users or data providers.
Value proposition	This service permits to upload your own list of species to use in the GBIF.ES portal or use a list uploaded by other GBIF.ES users or data providers.

Customer of the service/tool	https://www.gbif.es/wp-content/uploads/2017/05/gbif-logo.svg
User of the service/tool	Research use. This tool allows users to upload lists of species and work with that list within the Atlas. It is also possible to access lists made by others.
User Documentation	https://www.gbif.es/en/portal-nacional-de-datos/
Technical Documentation	https://www.gbif.es/en/portal-nacional-de-datos/
Product team	GBIF Spain, LifeWatch
License	https://www.gbif.es/en/politica-de-privacidad/
Source code	-
Testing	-

Service/Tool name	**GBIF Spain Collections**
Service/Tool url	https://colecciones.gbif.es
Service/Tool information page	https://colecciones.gbif.es
Description	GBIF.es data access under biogeographic context provides access with advanced facets to GBIF biodiversity data under a biogeographic context. Institutions, collections and datasets list, description and linked to occurrences.
Value proposition	This service allows users to search for data using DOIs. The database contains information on institutions, collections and species records. Currently there are only data from entities that collaborate with GBIF
Customer of the service/tool	https://www.gbif.es/wp-content/uploads/2017/05/gbif-logo.svg
User of the service/tool	The access to these collections is open, this service is aimed both for people who want to learn and for research. Serves as a record.
User Documentation	https://www.gbif.es/en/portal-nacional-de-datos/
Technical Documentation	https://www.gbif.es/en/portal-nacional-de-datos/
Product team	GBIF Spain, LifeWatch
License	https://www.gbif.es/en/politica-de-privacidad/
Source code	-
Testing	-

Service/Tool name	**GBIF Spain Regions**
Service/Tool url	https://regiones.gbif.es

Service/Tool information page	https://regiones.gbif.es
Description	GBIF.es data access under biogeographic context provides access with advanced facets to GBIF biodiversity data under a biogeographic context. Map tool to browse states, territories, local government areas and biogeographic regions.
Value proposition	Map tool to browse states, territories, local government areas and biogeographic regions.
Customer of the service/tool	https://www.gbif.es/wp-content/uploads/2017/05/gbif-logo.svg
User of the service/tool	This open tool shows, through a map, the species of the different regions of Spain. It is mainly oriented to research and academic use. It is also useful for any citizen who lives in an area of which this service has information.
User Documentation	https://www.gbif.es/en/portal-nacional-de-datos/
Technical Documentation	https://www.gbif.es/en/portal-nacional-de-datos/
Product team	GBIF Spain, LifeWatch
License	https://www.gbif.es/en/politica-de-privacidad/
Source code	-
Testing	-

Service/Tool name	**GBIF Spain Occurrences**
Service/Tool url	https://registros.gbif.es
Service/Tool information page	https://registros.gbif.es
Description	GBIF.es data access under biogeographic context provides access with advanced facets to GBIF biodiversity data under a biogeographic context. List of occurrence data, georeferenced records and data set search statistics. Allows access to the information of each record and the quality tests performed.
Value proposition	List of occurrence data, georeferenced records and data set search statistics. Allows access to the information of each record and the quality tests performed
Customer of the service/tool	https://www.gbif.es/wp-content/uploads/2017/05/gbif-logo.svg
User of the service/tool	Research use. Access to occurrence biodiversity data published by Spanish providers
User Documentation	https://www.gbif.es/en/portal-nacional-de-datos/
Technical Documentation	https://www.gbif.es/en/portal-nacional-de-datos/
Product team	GBIF Spain, LifeWatch

License	https://www.gbif.es/en/politica-de-privacidad/
Source code	-
Testing	-

Service/Tool name	**GBIF Spain Images**
Service/Tool url	https://imagenes.gbif.es
Service/Tool information page	https://imagenes.gbif.es
Description	GBIF.es data access under biogeographic context provides access with advanced facets to GBIF biodiversity data under a biogeographic context. Image storage and web services application. Images can be measured, calibrated, zoomed or downloaded
Value proposition	Image storage and web services application. Images can be measured, calibrated, zoomed or downloaded
Customer of the service/tool	https://www.gbif.es/wp-content/uploads/2017/05/gbif-logo.svg
User of the service/tool	Research and academic use. Visualize images found in the data portal for a taxon.
User Documentation	https://www.gbif.es/en/portal-nacional-de-datos/
Technical Documentation	https://www.gbif.es/en/portal-nacional-de-datos/
Product team	GBIF Spain, LifeWatch
License	https://www.gbif.es/en/politica-de-privacidad/
Source code	-
Testing	-

Service/Tool name	**GBIF Spain eLearning**
Service/Tool url	https://elearning.gbif.es
Service/Tool information page	https://elearning.gbif.es
Description	GBIF.es data access under biogeographic context provides access with advanced facets to GBIF biodiversity data under a biogeographic context. The GBIF.ES e-learning platform offers online workshops on data quality, data publication, data use, etc. open to whole GBIF community
Value proposition	The GBIF.ES e-learning platform offers online workshops on data quality, data publication, data use, etc. open to whole GBIF community

Customer of the service/tool	https://www.gbif.es/wp-content/uploads/2017/05/gbif-logo.svg
User of the service/tool	Platform offers online workshops open to whole GBIF community.
User Documentation	https://www.gbif.es/en/portal-nacional-de-datos/
Technical Documentation	https://www.gbif.es/en/portal-nacional-de-datos/
Product team	GBIF Spain, LifeWatch
License	https://www.gbif.es/en/politica-de-privacidad/
Source code	-
Testing	-

9.2 Initial ambition (in 2018)

The initial offer of LifeWatch about services that were going to be integrated into the EOSC marketplace is shown in the following table (see below).

From that extensive list, finally six services were selected to start working with. They were:

- 4-GBIF data access under biogeographic context; Observatories, Data Collections; JRU LW.ES, LW-PT, NGI-ES, NGI-PT [GBIF Spain and GBIF Portugal, LIP(Pt), IFCA(Sp)]; EGI
- 5-PAIRQURS (Life+ RESPIRA); Citizen Science, Data; JRU LW.ES, BSC [U. Navarra (Sp), BSC(Sp)]; EUDAT
- 6-Citizen Science Services; Citizen Science, Data Processing; JRU LW.ES, NGI-ES [BIFI(Sp), CREAF (Sp), GBIF node (Sp), IFCA (Sp), U.Granada (Sp)], U.Cordoba (Sp)]; EGI
- 7-Image Classification Deep Learning Tools; Citizen Science, Data Analysis; JRU LW.ES, NGI-ES [BIFI (Sp), IFCA (Sp)]
- 14-Digital Knowledge Preservation Framework; Virtual Lab, Data Management, Analysis Preservation; JRU LW.ES [IFCA (Sp), U.Sevilla(Sp), CITIC(Sp)]; EGI
- 15-Remote Monitoring and Smart Sensing; Virtual Lab, Data Analysis; JRU LW.ES [IFCA (Sp), U. Sevilla (Sp), Andalucia SmartCity (Sp)]

However, considering the original proposal, there were some changes that arose from causes outside of our control such as the decision of the service providers of not going on with the development or deployment of some e-services (e.g. PAIRQURS, CINDA). In addition, although all the services we have worked with were planned to be part of the LifeWatch ERIC catalog in the near term, after two years this action is still pending. As a consequence, we searched for other LifeWatch collaborators, and the final list of services we targeted to integrate in the marketplace was:

- Glacier Lagoons of Sierra Nevada ("Lagunas de Sierra Nevada")
- GBIF.es
- Plant classification

- Remote monitoring and smart sensing

Lastly, due to some issues related to the service level agreements (SLAs) between LifeWatch ERIC – The Service Providers – EOSC-hub, the activity of the USE in this project was delayed with regard to the initial plan. Therefore, the roadmap was modified in order to get the same level of development as the other service providers by December 2019 (M24).

Here it is important to remark that both services from IFCA-CSIC, Remote Monitoring and Smart Sensing and Plant Classification, are hosted thanks to EGI FedCloud resources under the LifeWatch Virtual Organization. IFCA provides computing resources via HPC, HTC and Cloud Computing. HTC is connected to the EGI infrastructure and part of the Cloud Computing resources support different Virtual Organizations under the EGI Federated Cloud umbrella.

For GBIF.es, all the computer infrastructure necessary to maintain the GBIF.ES Biodiversity Data Portal is provided by the Institute of Physics of Cantabria (IFCA), which is the one who coordinates the activities of the National Grid Initiative in Spain (ESNGI). The union of the infrastructures of all the European NGIs constitutes the European Grid Infrastructure (EGI).

Service Name	Main area	Main teams involved	Links to EGI, EUDAT, INDIGO
1-Collaborative platform for observatories	**Observatories** Data Management	LW-Be, LW-Gr [VLIZ (Be), HCMR (Gr)]	EGI
2-Modelling Water Masses	**Observatories** Modelling	JRU LW.ES, NGI-ES, NGI-IT [IFCA(Sp), Ecohydros(SME) (Sp), aDevice(SME)(Sp), CITIC (Sp), INFN(It)]	INDIGO, EGI
3-Data Services	**Observatories** Data Management	LW-Gr [HCMR (Gr)]	-
4-GBIF data access under biogeographic context	**Observatories** Data Collections	JRU LW.ES, LW-PT, NGI-ES, NGI-PT [GBIF Spain and GBIF Portugal, LIP(Pt), IFCA(Sp)]	EGI
5-PAIRQURS (Life+ RESPIRA)	**Citizen Science** Data	JRU LW.ES, BSC [U. Navarra (Sp), BSC(Sp)]	EUDAT
6-Citizen Science Services	**Citizen Science** Data Processing	JRU LW.ES, NGI-ES [BIFI(sp), CREAF (Sp), GBIF node (Sp), IFCA (Sp), U.Granada (Sp)], U.Cordoba (Sp)].	EGI
7-Image Classification Deep Learning Tools	**Citizen Science** Data Analysis	JRU LW.ES, NGI-ES [BIFI (Sp), IFCA (Sp)]	EGI
8-Genetic Services	**Virtual Lab**	LW-Gr [HCMR (Gr)]	-
9-MiroCT	**Virtual Lab**	LW-GR [HCMR(Gr)]	-
10-R Services	**Virtual Lab** Analysis	LW-Gr, LW-Be, JRU LW.ES [HCMR (Gr), VLIZ (Be), IFCA (Sp)]	EGI
11-Semantic Tools	**Virtual Lab** Data	LW-IT [UniSalento (It)/INFN]	-
12-Phytoplankton VRE	**Virtual Lab** Data Analysis	LW-IT [UniSalento (It)/INFN]	-
13-Ecological Data analysis platform	**Virtual Lab** Data Analysis	LW-IT [UniSalento (It)/INFN]	-
14-Digital Knowledge Preservation Framework	**Virtual Lab** Data Management Analysis Preservation	JRU LW.ES [IFCA (Sp), U.Sevilla(Sp), CITIC(Sp)]	EGI
15-Remote Monitoring and Smart Sensing	**Virtual Lab** Data Analysis	JRU LW.ES [IFCA (Sp), U. Sevilla (Sp), Andalucia SmartCity (Sp)]	EGI
16-TRUFA	**Workflows**	JRU LW.ES [IFCA (Sp)]	INDIGO, EGI
16+1* -Declic[3]	**Workflows**	NGI-FR, +NGI-ES [PGTB(Fr), UPV(Sp), France-Grilles]	EGI

9.3 Final software architecture and integration

Not applicable to LifeWatch services.

With the exception of Remote Monitoring and Smart Sensing, all the thematic services integrated are available on the website of their corresponding providers.

Besides registration in the EOSC Portal and Marketplace, there was not any integration performed between LifeWatch and EOSC-hub services as part of the project even if most of the services (as from the previous table) were already integrated with EGI, EUDAT or INDIGO services.

9.4 Impact and exploitation

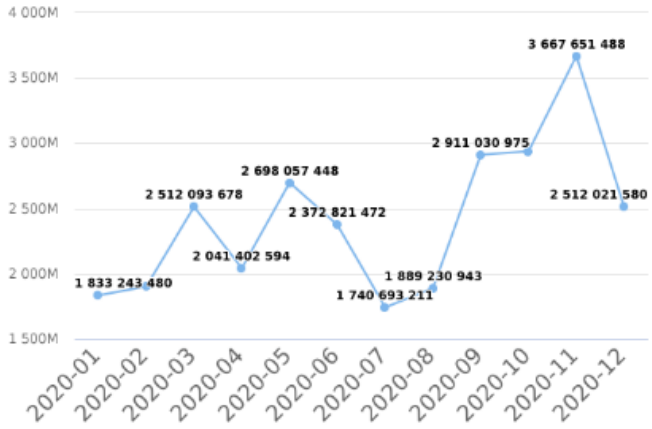
Although a deeper analysis is needed on the side of our service providers, in general terms we can state that the integration of the services in the marketplace has been really beneficial to gain visibility and boost the number of new users.

Considering the IFCA-CSIC service for plant classification (https://marketplace.eosc-portal.eu/services/plant-classification), the number of users increased 116.1% and the number of sessions 64.4%. Regarding the metrics of Google analytics for this service (available from June 2020), the number of users has increased about 88% and the sessions 71.7%.

In the case of the Remote Monitoring service, also from IFCA-CSIC, we do not have numbers to compare with, since the service was released for the first time in June 2020.

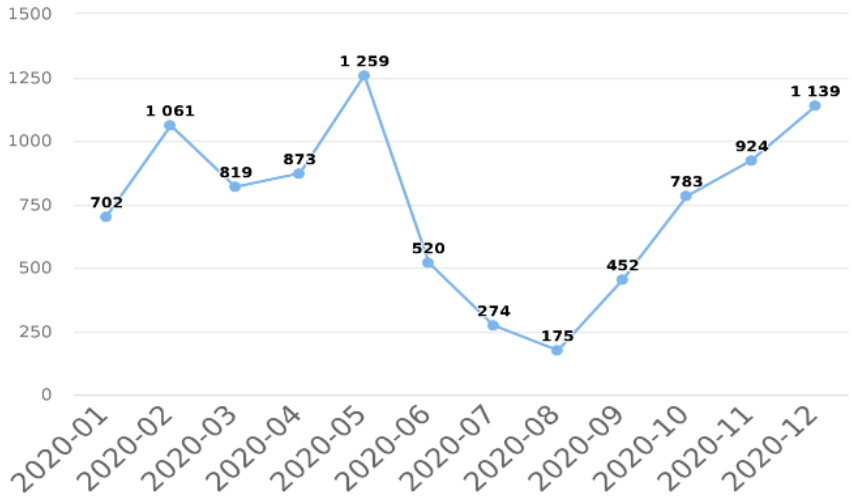
From the official activity report from GBIF Spain, generated in January 2021 and available here, we have extracted the following metrics:

Requests of datasets around the world with origin in Spain:



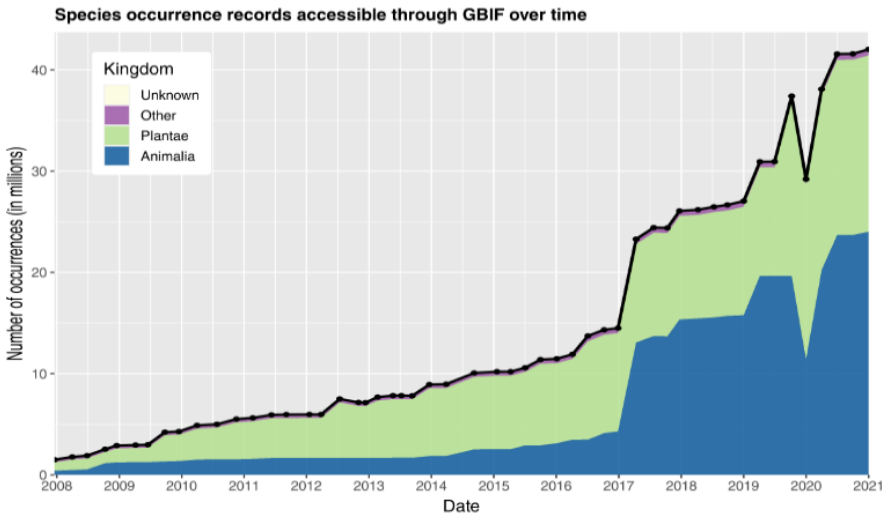
Number of occurrence records downloaded via GBIF.org published by institutions in Spain

And the national demand of records was the following:



Monthly downloads requested by users in Spain

In parallel we have detected an increase in the number of records published by Spanish institutions.



Number of records published by institutions in Spain, categorized by kingdom

As evidence, GBIF Portugal, which works closely with GBIF.ES (Spain), has demanded to include its services into the EOSC marketplace to get the same advantages.

Additional results and detailed numbers are being provided in this respect in the framework of the WP13.

9.5 Lesson learnt

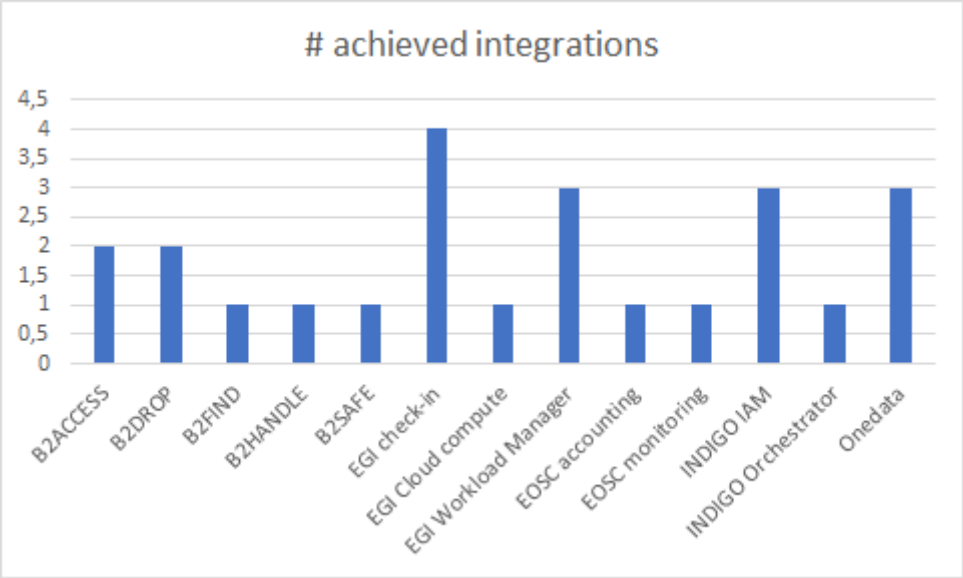
In EOSC-hub, the University of Sevilla (USE) has been the organization in charge of managing the integration of the e-services linked to LifeWatch into the Marketplace. In this sense, it has been difficult sometimes to be such an intermediate between LifeWatch and EOSC. For example, during the negotiations of the terms and conditions of the SLA. In this period, the integration process was

paralysed. The SLA needed to be agreed upon by the service providers (their services were not, and are not today, part of the LifeWatch catalogue), LifeWatch and EOSC-hub. In the meantime, the University of Sevilla (the partner of EOSC-hub) had to deal with the service providers to continue with the integration tasks without delaying them too much regarding the planned schedule. At this stage, there was a turning point. We agreed on a new roadmap with EOSC and started working with other service providers. Moreover, the issue of SLA was successfully solved. Finally, we achieved to incorporate more than 8 services in the marketplace.

Even if there is a happy ending, bearing in mind all the integration process, we consider that it would be more straightforward if the service providers are the contact persons instead of another organisation. I would recommend simplifying the communication channels between EOSC and the service providers.

10 Conclusions

As reported in the previous section the activities of the Thematic services were really successful as all TSs are reporting an increase of their user base or engagement with respective communities. All TS have exposed their services via the EOSC marketplace: 30 services in total and 24 among EOSC services have been achieved and summarised here below:



In general the availability of core services and resources in EOSC has proven to be essential for the further enhancement of the thematic services and all TSs report a positive impact on their respective communities.

However some common points on the lessons learn might be considered for the future of EOSC:

- The use of the Marketplace concept for services that are free for use or open access might not be ideal for users (CLARIN, DODAS, WeNMR);
- The complexity of integrating new services requires a deeper work between thematic services providers and core service personnel and it should be considered that integration is not a one-time activity (OpenCoastS, DODAS, DARIAH);
- Effort needed for onboarding in the EOSC marketplace has been sometimes underestimated (DARIAH, LifeWatch);
- Training and dissemination events play an important role in engaging with more users and expanding beyond the existing user base (DODAS, ECAS, WeNMR).

Each Thematics service, as detailed also in D3.5, has its own path for exploitation and future sustainability in large part via future EC funded projects.

11 References

No	Description/Link
R1	D7.1 “First Thematic Service software release”, https://documents.egi.eu/document/3411
R2	D7.2 “First report on Thematic Service architecture and software integration”, https://documents.egi.eu/document/3412
R3	D7.3 ”First report on Thematic Service exploitation”, https://documents.egi.eu/document/3577
R4	Oliveira, A., A.B. Fortunato, J. Rogeiro, J. Teixeira, A. Azevedo, L. Lavaud, X. Bertin, J. Gomes, M. David, J. Pina, M. Rodrigues, P. Lopes, 2020. OPENCoastS: An open-access service for the automatic generation of coastal forecast systems, Environmental Modelling and Software, 124: 104585. DOI: 10.1016/j.envsoft.2019.104585
R5	Oliveira, A.; Rodrigues, M.; Rogeiro, J.; Fortunato, A. B.; Teixeira, J.; Azevedo, A.; Lopes, P. 2019, Opencoasts: an open-access app for sharing coastal prediction information for management and recreation, 794-807pp, Computational Science-Lecture Notes-ICCS2019, Vol. 11540, Computational Science-Lecture Notes-ICCS2019
R6	Marta Rodrigues, João Rogeiro, S. Bernardo, A. Oliveira, A. B. Fortunato, J. Teixeira, Pedro Lopes, A. Azevedo, J. Gomes, M. David, J. Pina, 2019.OPENCoastS: an operational, on-demand forecast tool to support coastal management, Proceedings of the 14th International MECOAST Congress on Coastal and Marine Sciences, Engineering, Management and Conservation MEDCOAST 2019, E. Ozhan (Editor), vol. 1, 57-67.
R7	D7.4 ”Second Thematic Service software release”, https://documents.egi.eu/document/3642