D8.2 Final report on Competence Centre key results and exploitation status and plans

Lead Partner:	EGI Foundation
Version:	1
Status:	FINAL
Dissemination Level:	Public
Document Link:	https://documents.egi.eu/document/3631

Deliverable Abstract

WP8 includes eight Competence Centres (CCs) that work on establishing infrastructures to support users in coping with the data deluge in various compute intensive data analysis scenarios. This document provides the final report about the CC activities and informs readers about the use cases the CCs worked on, the implementations they reached in the past 3 years with the use of EOSC-hub and other EOSC-related services. The document also covers the impact of this work on the various research infrastructures and scientific communities that are linked to the CCs, and outlines future work for after EOSC-hub, based on the reported achievements.

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:	Gergely Sipos	EGI Foundation/WP8	01/04/2021
Moderated by:	Malgorzata Krakowian	EGI Foundation/WP1	
Reviewed by:	Debora Testi Gianni Dalla Torre	CINEC/WP7 EGI Foundation/WP11	24/02/2021 22/02/2021
Approved by:	AMB		

DOCUMENT LOG

Issue	***Date***	***Comment***	***Author***
v.0.1	15/09/2020	Table of content	Gergely Sipos (EGI Foundation), AMB, CC coordinators
v.0.2	09/12/2020	Full draft for WP review	Gergely Sipos (EGI Foundation), Justin Clark-Casey (EMBL-EBI), Michal Prochazka (Masaryk University), Andrew Lahiff (UKAEA), Marcin Plociennik (PSNC), Thierry Carval (Ifremer), Peter Thijsse (MARIS), Carl-Fredrik Enell, Rikard Slapak (EISCAT Association), Luca Trani (KNMI), Javier Quinteros (GFZ-Potsdam), Hanno Holties (ASTRON), Alex Vermeulen (ICOS ERIC), Peterseil Johannes (Umweltbundesamt GmbH), Eric Yen (Academia Sinica)
v.0.3	19/02/2021	Full version for external review	Gergely Sipos (EGI Foundation)
v.0.4	31/03/2021	Update based on review feedback	Several authors (see document issue v.0.2)
v.1	01/04/2021	Final	Gergely Sipos (EGI Foundation)

TERMINOLOGY

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

Contents

Executive summary

WP8 includes eight Competence Centres (CCs) that work on establishing infrastructures to support scientific communities in coping with the data deluge in various compute intensive data analysis scenarios. Each CC operates as a project on its own, with a small consortium composed of representative institutes from Research Infrastructures, experts of relevant e-infrastructure services, and software/technology developers. CCs expect to bring scalable and sustainable service setups for ELIXIR, Fusion (ITER), Euro-Argo & SeaDataNet[1], EISCAT_3D, EPOS-ORFEUS, LOFAR (SKA), ICOS & eLTER[1] and Disaster Mitigation. The overall objective of the CCs is to co-design, co-develop and validate services for these communities by mobilising generic services from the so called 'EOSC-hub common services' portfolio provided by e-infrastructures.

The work started within each CC drafting plans concerning their envisaged use of the EOSC-hub common services, based on experiences from predecessor projects (primarily EGI-Engage, EUDAT2020, EOSCpilot). These plans were shared across the CCs during the kick-off meeting, during the EOSC-hub week, and the monthly CC coordinators teleconferences. By the beginning of 2019, the plans evolved into documents within the EOSC-hub Wiki and into technical requirements in EOSC-hub JIRA, where those were directly communicated to service developers outside WP8. The Wiki pages and the technical requirements together formed M8.1 in March 2019).

During the 3 years the 10 communities experimented with 15 common services from the EOSC-hub catalogue, and with 6 services/technologies from outside the project (See Appendix 1 for a table overview of these across the CCs). After successful validation and integration, 10 of the EOSC-hub common services progressed into production setups within the communities.

3 CCs reached mature integration between common services and their community-specific services and opened up their setups for external users via the EOSC Portal: Argo (ARGO floats data discovery), EISCAT_3D (EISCAT Data Access Portal), Fusion (PROMINENCE).

ELIXIR interfaced the ELIXIR AAI with AAI proxies from EOSC-hub and registered 3 institutional clouds in EOSC Portal (CESNET, CSC, EBI).

One additional community reached mature setup, and at the time of writing its service is in the onboarding queue in the EOSC Portal: ICOS (ICOS Portal).

Two communities are planning to stabilise their setup in 2021 and then onboard it in the EOSC Portal: Radioastronomy (LOFAR Science Products) and Disaster Mitigation (iCOMCOT Portal). The extra development required to reach the service and then to operate it in EOSC will be covered by the EGI-ACE project for ASTRON (and for SURFsara as its technical partner) and AS.

All the other CCs resulted in 'validated solutions', together with known limitations, and with unexplored validation options. These findings are expected to feed into future projects linked to the respective communities.

[1] The Euro-Argo and SeaDataNet communities are in a single CC. The ICOS and eLTER communities are in a single CC. This is why the 8 CCs include 10 communities and Section 3 has 10 sub-sections.

Section 1 provides further background information about the CCs and their related research communities.

Section 2 of the document details the work performed by the 10 communities. Every section follows the same structure, describing the initial ambition of the community (from 2018), describes the progress and key results achieved, captures the lessons learnt, summarising the impact of the work, and providing plans beyond EOSC-hub.

Section 3 highlights the common needs and recurring success stories and the main lesson learnt from the CC activity, capturing those in 10 points.

1 Introduction to Competence Centres

The EOSC-hub Work Package 8 is composed of 8 Competence Centres (CCs). Each CC fosters the use of advanced digital capabilities and services of EOSC-hub within early adopter research communities, supporting them in data- and computing-intensive science. CCs are driven by well-established and mature research infrastructures from ESFRI or by international scientific collaborations that require integrated data and computing services from multiple European e-infrastructure providers, typically from EGI, EUDAT and GÉANT. CCs operate independently from each other but share interest and needs in using common solutions from the EOSC-hub service catalogue to setup community-specific services that can expand EOSC with science discipline specific capabilities.

CCs were selected during the EOSC-hub proposal preparation in 2017 with an open process that invited research communities, research infrastructures, software developer teams and e-infrastructure service providers to form small, focussed piloting projects that evaluate generic services from European e-infrastructures for the support of international, scientific communities. The proposal editorial team selected the best 8 from the 48 (!) submissions, and included them as Competence Centres in the project:

1. ELIXIR CC (linking to the ELIXIR ESFRI Landmark)
2. Fusion CC (linking to the EUROfusion programme)
3. Marine CC (linking to the Euro Argo ESFRI landmark and to the SeaDataNet community)
4. EISCAT_3D CC (linking to the EISCAT_3D ESFRI landmark)
5. EPOS-ORFEUS CC (linking to the EPOS ESFRI landmark)
6. Radioastronomy CC (linking to the LOFAR pathfinder of the Square Kilometer Array ESFRI landmark)
7. ICOS-eLTER CC (linking to the ICOS ESFRI landmark and to the eLTER ESFRI project)
8. Disaster Mitigation CC (linking to regional initiatives in the Asia-Pacific region)

Several of these CCs are the continuation of similar piloting activities that were part of the EGI-Engage and EUDAT2020 projects between 2015-2017.

The CCs integrate community-specific data repositories, applications, portals, AAI system (Authentication-Authorization Infrastructures) with the generic services[2] offered by EOSC-hub with the ultimate goal to provide scalable and robust data management and processing services for researchers within their own domains. The average effort level is approximately 1.6 Full Time Equivalent per CC, distributed among 3-5 institutes (depending on the CC). Every CC runs through the whole EOSC-hub project for 3 years.

[2] EOSC-hub generic services cover the following areas: Data discovery and access; Federated compute; Processing and orchestration; Data and metadata management; Data preservation; Sensitive data; Identification-authentication-authorization and attribute management; Marketplace and order management; Integrated operations support systems; Monitoring-accounting-messaging-security tools; Helpdesk; Application store, software repositories.

2 Key results, exploitation, plans

2.1 ELIXIR

2.1.1 Initial ambition (in 2018)

The CC will enable ELIXIR to establish an ELIXIR Compute Platform (ECP) which allows ELIXIR cloud and data providers to share cloud compute and storage capacity to replicate and share reference datasets with each other and with their users. The platform aims to enable researchers to combine technical components of the ELIXIR Compute Platform services into a seamless ecosystem, thereby creating a science ready, standardised interface to the key resources and technological capabilities that are available for life sciences. The ECP aims to leverage the EOSC Service Catalogue to enable two related yet distinct activities for ELIXIR:

1. establish a federation of cloud sites, each providing storage and compute capacity for researchers, connected to a Reference Data Set Distribution Service (RDSDS) that enables the staging of 'ELIXIR Core Data Resources' to the cloud sites on-demand;
2. Establish a ‘container replication and orchestration service’ that enables application providers to deploy containerised community/reference applications to any of the federated cloud sites.

2.1.2 Progress and key results

2.1.2.1 Reference Data Set Distribution Service

During the first half of the project, the ELIXIR CC developed and deployed the first version of RDSDS to ELIXIR members CSC and CESNET, making available 1000 Genomes and MMGELIXIR reference datasets. More datasets were planned but no work had started to connect consuming systems. Moreover, due to personnel changes and unexpected complexity of implementation no work had started on the second use case to establish a container replication and orchestration service.

Very soon after that demonstration the life sciences data analysis landscape started to shift. The Global Alliance For Genomics and Health (GA4GH), a recently established organization dedicated to creating common standards for sharing and processing genomic data, started to release APIs for implementing a distributed analysis framework. These APIs aim to standardize both the means of locating data through the Dataset Repository Service API (DRS) but also the invocation of workflows (through the Workflow Execution Service API - WES) and the locating of containerised tools (through the Tools Repository Service API - TRS). GA4GH also initiated work to control the sharing of sensitive data (GA4GH passports).

We realized that these APIs would be an excellent means of fulfilling the second use case: Other ELIXIR nodes and EMBL-EBI could work on creating or adapting existing infrastructures to provide WES and TRS implementations. RDSDS could become a DRS implementation as it was already indexing data locations. By adopting the GA4GH APIs, components could become plug and play with implementations written elsewhere to create a distributed analysis system. For example, a WES implementation written by the NIH could pull data from RDSDS via its DRS implementation, without having to learn each individual resource's way of providing for data download.

It quickly became apparent that the coding of the first version of RDSDS would not form a solid basis for this adaptation. Therefore, we took the decision to rewrite RDSDS. Today, this is deployed in ELIXIR partners at CSC and CESNET and has been used in GA4GH demonstrators. RDSDS is now a GA4GH DRS implementation that is able to take part in distributed workflows through GA4GH compliant systems. As a DRS implementation it also helps explore the practical implementation of these systems and further refinement of these APIs. In the October 2020 GA4GH Plenary, RDSDS was successfully used as part of the ELIXIR 'stack' in the GA4GH connections session. A stack is an implementation of GA4GH APIs (DRS, WES, TRS and others) that can run a distributed workflow from start to finish. ELIXIR is not the only stack that exists - there is also one based on the Broad Institute's Terra system as well as stacks that come from commercial vendors, such as SevenBridges and DNAStack. As these stacks mature it should be possible to mix and match their components (e.g. use RDSDS as the DRS in the Terra stack).

At the same time, other developments started to happen in the data distribution space. For example, the OneData system has started to emerge. This allows both pre-caching of datasets at compute clouds, and downloading of chunks of data into clouds only when and where they are needed. The cached datasets are then available to analysis workflows. Judging by the interest from other ELIXIR nodes and in other research and e-infrastructures, this will be a more popular approach to dataset distribution than the RDSDS proposal. Therefore, RDSDS will go forward chiefly as a DRS implementation, providing ELIXIR AAI (later LS AAI)/GA4GH passport data location services, and providing a Globus data transfer option. Pull data distribution (as opposed to push) will be left to systems such as OneData.

As part of this change in direction, ELIXIR did not move towards federated cloud provisioning across its members. Rather, it is supporting the registration of members as individual providers in EOSC at this time. For example, both EMBL-EBI and de.NBI are shown as compute providers in the EOSC marketplace.

Figure 1 shows an illustrative diagram of this arrangement. RDSDS is shown in the core not necessarily because it will be provided as an EOSC core service, but rather that it plays the same role of providing information on how to discover datasets. In the first step, RDSDS indexes public repositories, such as the Metabolights and PRIDE repositories maintained by EMBL-EBI. In the second step, some community specific analysis workflow accesses its own experimental data held in EOSC. In step three, that workflow queries RDSDS to discover how to access public datasets that it needs to perform an integrated analysis. Lastly, the workflow accesses those datasets and performs the analysis.

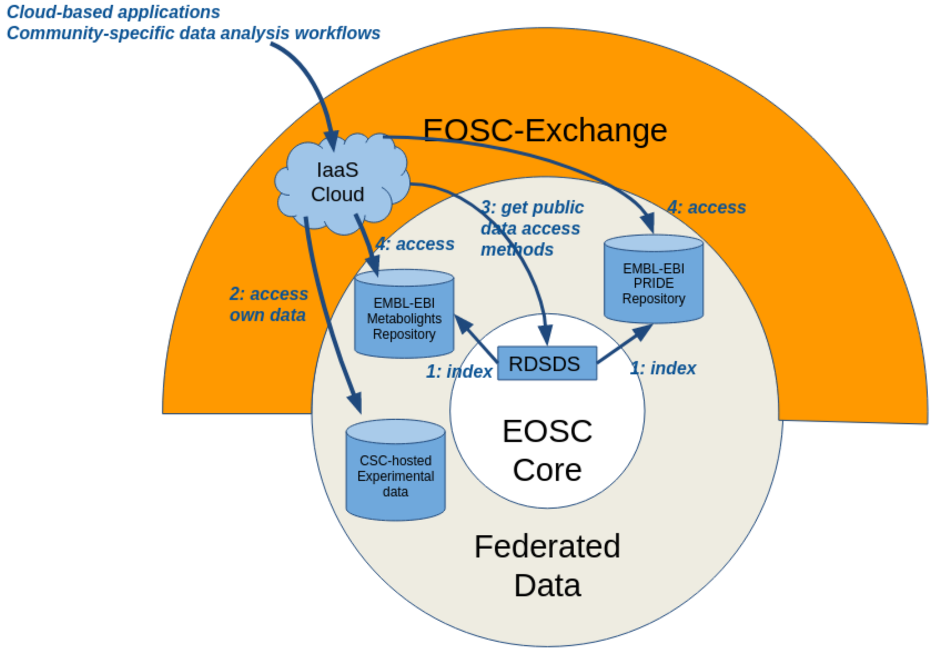

Cloud-based applications
Community-specific data analysis workflows
EOSC-Exchange
IaaS Cloud
3: get public data access methods
4: access
4: access
2: access own data
EMBL-EBI Metabolights Repository
EMBL-EBI PRIDE Repository
RDSDS
1: index
1: index
CSC-hosted Experimental data
EOSC Core
Federated Data

Fig.1 – Dataset Distribution Service

2.1.2.2 Life Sciences AAI

In parallel with the RDSDS effort, significant work has taken place in the ELIXIR Competence Centre to establish a Life Sciences AAI (LS AAI) that will greatly expand upon the pre-existing ELIXIR AAI. ELIXIR Competence Center participated in the definition of the migration plan from existing ELIXIR AAI into the Life Sciences AAI because all services and identity providers connected to ELIXIR AAI have to be migrated, that means also EOSC AAI components like EGI Check-in and EUDAT B2ACCESS. The plan is not yet fully implemented because the Life Sciences AAI is not fully operational yet. ELIXIR Competence Center also provided input into the definition of the final shape of Life Sciences AAI materialised in the 'EOSC-Life Access and user management system for life science – the blueprint'[3] publication.

Using the ELIXIR AAI statistics page[4] we can observe that users are using ELIXIR identities to access EGI Check-in and B2ACCESS then reach EOSC services. Until now, we can see that users from 9 institutions have accessed EOSC services using their ELIXIR identity through connected EGI Check-in and EUDAT B2ACCESS gateways.

2.1.2.3 Cost model for cloud provisioning

As part of the ELIXIR Competence Centre, we also looked at producing a common ELIXIR cloud provisioning costing model to inform future virtual access models. Ultimately, this proved too difficult to achieve without knowing which virtual access model or models would be used in EOSC. For example, there is the question of what unit of consumption would be used - CSC has the concept of a billing unit that accrues at different rates to different services (compute, storage, etc.) whilst

[3] https://doi.org/10.5281/zenodo.3386307

[4] https://login.elixir-czech.org/statistics

EMBL-EBI has no comparable model as costs are estimated separately per collaborator. It would be a very high effort to create a similar billing unit for EMBL-EBI. This effort is difficult to justify if the reimbursement model is not known upfront, particularly in the light of difficulties getting reimbursed for current cloud costs in existing EC grants. We find that the recently released EOSC-hub briefing paper on the provision of cross-border services[5] includes a very good discussion of the difficulties that need to be tackled in coming up with a cross-border reimbursement model.

2.1.3 Lessons learnt

- In a scientific environment of active development and high complexity, it is important to have code that is modular and highly adaptable. We learnt this with RDSDS - the first version was monolithic and proved difficult to adapt to the DRS API. The current version is much more flexible and written with the cloud in mind. For example, the indexing will be Kubernetes native and not tied to any particular software cluster environment.
- Pre-caching of datasets at compute clouds and downloading of chunks of data into clouds only when and where they are needed is the preferred way of data distribution. A central data catalogue with location service, and distributed data caching services at each cloud (e.g. OneData) could be the infrastructure to reach such a setup.
- In the AAI arena we discovered that from a user’s point of view it is complicated to authenticate with services, not technically but conceptually. Users have to be aware of the affiliation with their research infrastructure and their role they would like to have at the service. For example, when accessing an EOSC service behind EGI Check-in, the user needs to select the ELIXIR research infrastructure as the entity for authenticating the user. However, on the EGI Check-in page there are also particular institutions listed that may include the user’s institution. In this case, the user can be confused because he/she does not know whether to select ELIXIR or his/her home organization. IdP hinting[6] could help in such situations by letting service providers pre-select authentication providers.
- Cloud providers work with different billing units and cost models within ELIXIR. It would be a very high effort to bring all such providers to a common billing unit to enable cross-site billing (with virtual or real funding). This effort is difficult to justify if the reimbursement model is not known upfront.

2.1.4 Impact

The CC contributed to the advancement of the ELIXIR AAI to become a Life Science AAI, serving the whole life and biological sciences domain. The LS AAI is expected for production opening in 2021.

The technical developments, demonstrations and discussions in the other areas of the CC (data staging and cloud cost models) helped the ELIXIR community advance its understanding about

[5] EOSC-hub Briefing Paper – Provision of CrossBorder Services: https://www.eosc-hub.eu/sites/default/files/EOSC-hub%20Briefing%20Paper%20-%20Provision%20of%20Cross-Border%20Services%20-%20final_0.pdf

[6] A specification for IdP hinting (by the AARC project): https://aarc-project.eu/wp-content/uploads/2019/04/AARC-G049-A_specification_for_IdP_hinting-v6.pdf

possibilities, about community preferences. The lessons learnt will feed into the future planning of infrastructure development activities of ELIXIR.

2.1.5 Future plans beyond EOSC-hub

For RDSDS, beyond EOSC-hub we will continue to work within ELIXIR to explore its direct use in various projects and use it as a means to help other ELIXIR data resources adopt the DRS API. The CC also suggests to ELIXIR to include user-driven, on-demand data staging tools in its portfolio, and first share OneData experiences within its compute platform in 2021-22, then conduct broader deployment and tests.

ELIXIR AAI will be migrated to the LifeScience AAI which is operated by e-Infrastructures and funded by EOSC-Life. We have defined a migration plan that will move all registered services from ELIXIR AAI into the LifeScience AAI including EGI Check-in and EUDAT B2ACCESS. There will be just minor changes from the user point of view (only logos and web page layout).

We also implemented support for GA4GH Passports into the ELIXIR AAI and now we are working on supporting GA4GH in the LifeScience AAI. When it is done, all EOSC services will be able to utilize GA4GH Passport for managing access to sensitive data.

2.3 Fusion

2.3.1 Initial ambition (in 2018)

The CC's ambition is to assess whether the services provided by EOSC are suitable for use cases within the fusion community. This work has been split into two; one storage specific and one compute specific. The reason behind these investigations is in preparation for ITER data handling and analysis, which represents a major technological challenge for the fusion community increasing the volume of output by three orders of magnitude from current experiments.

For storage we wish to investigate replication between sites. Since ITER is an international experiment it is likely there will be at most two European sites which will host a fraction of the data and some portion of that data will need to be readily available for analysis at several centres of excellence. Other sites may wish to access the data, but in this case the analysis is not time critical and so are not considered here. It is envisaged that automated replication of data will be key to this work but will require the underlying technology to support high speed IO and replication. As this is primarily a technology assessment, we have specifically omitted security implications. However, if a suitable EOSC technology is identified then this will need to be taken into consideration for final usage.

In compute terms, the CC is again being driven by the needs of ITER. It is not anticipated that one single site will be able to meet the needs of ITER data analysis and, indeed, pre-testing. One partner has already demonstrated a service which allows modelling code to be run at any site with available (and suitable) compute resources. This work needs to be extended to support ITER type operations; specifically the execution of full workflows. As such we are taking a twofold approach; running an existing 'real life' use case from the MAST tokamak, taking raw output from one of the diagnostic tools and processing it to science products, and also using the ITER Integrated Modelling and Analysis Suite (IMAS) to test prototypical ITER workflows. The idea is that making use of cloud resources will better allow sites to process 'intershot' data at a scalable level and maintain a smaller ecological footprint than would otherwise be necessary.

2.3.2 Progress and key results

2.3.2.1 Storage technology evaluations

Technology evaluations of B2SAFE and OneData were carried out as they are both storage systems available in EOSC able to support replication and distributed data access. The evaluation criteria included

- supporting replication to defined sites,
- access via standard protocols (POSIX, HTTP, S3 or SWIFT),
- automated integrity checking with self-healing,
- provision of globally unique and resolvable persistent identifiers,
- ability to provide the “best” replica (e.g. geographically closest or lowest latency) and the ability to attach metadata.

For B2SAFE the plan was to make use of existing deployments at STFC, CINECA and PSNC, with data first written to STFC then replicated to CINECA and PSNC. STFC was chosen as it is physically close to CCFE. CINECA hosts the MARCONI/Fusion HPC facility and so it would be beneficial for fusion data to be hosted there due to the proximity of fusion compute resources. Getting access to the three B2SAFE instances was very difficult and took over 1.5 years (!). Replication from STFC to CINECA was eventually configured and successfully tested but technical issues at PSNC prevented data from being replicated there.

There were several challenges to data access. When data is replicated across multiple B2SAFE instances there is no existing way for automatically finding the nearest accessible replica, or even just a functional replica. Also, trying to access data from different instances requires not just knowing the hostname of the storage system but also the path to the data, which is different at different sites. Furthermore, the three B2SAFE instances support different access protocols. STFC supports only the HTTP API while CINECA supports WebDAV. All three sites support iRODs however we require a standard protocol.

For OneData the initial test was to check if OneData can satisfy the goal of enabling a laptop connected to the internet to access remote data, and furthermore test the replication functionality provided by OneData from a remote provider to a local one. The tests made use of an existing Onezone deployed at PSNC. A space was set up at PSNC supported by a provider at PSNC and a new provider deployed at IRFM. The provider at PSNC was populated with experimental data from WEST, the tokamak experiment operated by IRFM in France. This data was replicated to the IRFM provider. The tests involved running a small Python script which reads in the “magnetics” data structure from the data files using the IMAS data access layer. The tests were carried out on a Linux laptop at IRFM. As expected, reading data from the local IRFM OneData provider was faster than reading from the remote PSNC provider, in fact five times faster. More details on the OneData tests carried out on WEST can be found in our interim report[7].

One limitation found with OneData is that there are strict version requirements between the OneData provider and client which can be problematic. The default installation procedure for Oneclient installs the latest version which might not be compatible with the Oneprovider.

As with B2SAFE, OneData does not provide the ability to automatically find the nearest provider when multiple providers host the data.

2.3.2.2 Integration of UDA and B2SAFE

The integration of B2SAFE with the Unified Data Access (UDA) service, which provides a standard API for data access, was investigated. UDA is used as the data access layer for an increasing number of fusion codes. Integrating UDA with B2SAFE would ensure that input data will be accessible to fusion codes regardless of data locality.

A UDA server typically accesses the underlying data files from a standard network-attached filesystem, but there is also an experimental plugin for accessing data from AWS S3. This works by copying the files from S3 to a temporary cache on the local disk. Two different implementations of

[7] Data Replication and Access Testing using Onedata: https://documents.egi.eu/document/3560

an integration with B2SAFE were considered: one method employed B2SAFE mounted as a filesystem using webdavfs, thus requiring no modifications at all to UDA, and another using the B2SAFE HTTP API by using the existing S3 plugin as a starting point.

Testing made use of data from MAST, a spherical tokamak at CCFE. Initial tests were carried out by reading a variable from a MAST NetCDF data file via UDA in the UK and comparing performance where the source file is local and on B2SAFE at CINECA. Webdavfs includes a cache, and when files were already cached the access time compared to reading a file locally was very similar, as expected. When files were not cached B2SAFE was slower than the local disk, with the slowdown dependent on the file size. For a 3 MB file the time increased by a factor of 7, and for a 36 MB file the time increased by a factor of 33.

However, real applications of course do a lot more than just reading a single variable from a file. Therefore a more realistic test was carried out next by running EFIT++ to calculate the plasma equilibrium for the MAST tokamak, with data accessed from UDA with the files local to the UDA server or on B2SAFE at CINECA. The UDA server and machine running EFIT++ were again in the UK. When reading from B2SAFE, with files not cached, the difference in runtimes was negligible and turned out to be a few seconds difference compared to a total runtime of around 4 minutes. This test required 5 input data files with a combined size of 86 MB with individual files sizes ranging from 3 MB to 52 MB. The runtimes between B2SAFE cached, B2SAFE not cached and local files were all very similar since the data volumes are very small and the job is mainly CPU bound.

2.3.2.3 Choice of compute platform

We aimed to demonstrate making use of EOSC computational resources for running containerised modelling applications. This requirement derives from the fact that local resources are not scaled for peak demand and we wish to use the infrastructure provided by EOSC (and public cloud providers) as a scalable, non-vendor specific resource. This is an opportunistic use case where we wish to make use of any spare resources across multiple systems. Different parts of the workflows may have different resource requirements, ranging from a single core through to many cores/multi-nodes, so it is desirable that only the required resources are provisioned on each cloud as needed. If the workflow is running across multiple clouds each step of the workflow will need to access file(s) generated by previous steps, as necessary.

The Dynamic On-Demand Analysis Service (DODAS), which leverages the services developed in INDIGO-DataCloud, was initially evaluated as a platform for running processing jobs. It was found to be unsuitable for our use case as it would only support creating a batch system on-demand on a single cloud whereas we wanted the ability to run across multiple clouds.

We therefore made use of the PROMINENCE platform, developed in the EOSCpilot Fusion Science Demonstrator, which was explicitly designed for running jobs opportunistically across multiple clouds. We extended PROMINENCE to support automatically mounting OneData or WebDAV (e.g. B2SAFE) in jobs to allow data to be transparently accessed irrespective of location. The OneData support was initially tested with the OneData provider at PSNC, and WebDAV support was tested with B2SAFE at CINECA. One problem we experienced with the OneData integration is handling the strict version matching requirements between the client and server. A solution to this was found by having a variety of different versions of Oneclient pre-installed and available to jobs. The configuration endpoint of the Oneprovider is queried in order to find out which version of Oneclient to use. This enables PROMINENCE to run jobs requiring different versions of OneData to be supported transparently to users. We also extended PROMINENCE to support running Directed Acyclic Graph (DAG) workflows. As PROMINENCE internally uses HTCondor we were able to leverage HTCondor's DAGMan for running DAG workflows.

2.3.2.4 Choice of the AAI technologies

Available AAI technologies and their possible appliance to the Fusion community have been investigated. The concept presented in EOSC-hub meetings about Community AAI Proxies is fitting into the EUROfusion concept, so we considered the available proxies for our evaluation: B2Access, eduTeams, Indigo IAM, EGI Check-in. Different aspects have been taken into account including supported features, operational costs, maturity, sustainability, support.

In the first step we collected requirements in the fusion community. Then we checked and confirmed that all 4 solutions had the required technical features. We used the EOSC-hub AAI documentation[8] for the feature comparison.

As the next step two of the solutions have been tried out in the community: Indigo IAM and Unity that is part of B2Access. The tests were done in the EUROfusion AAI mini project including those partners where the proxy service was expected to be installed in-house inside the community. Due to encountered issues during the installation and tests, and after considering the effort and expertise that would be needed for the in-house proxy maintenance and operation, it was decided to move towards using "AAI proxy as a service" solution.

In the SaaS model we considered available support behind the services, expected sustainability in the long term and this assessment resulted in shortlisting eduTeams and EGI Check-in for deeper analysis. As the requirement was to have a dedicated instance of the proxy for the fusion community (and not to be shared with other communities), we checked/negotiated with the service providers the costs and terms of usage of dedicated instances possibly with SLAs.

After comparing the received offers we recommended eduTeams as community AAI to the EUROfusion AAI mini project. (The recommendation was presented to the community.) The mini project had a goal of piloting a selected solution and of setting up the community AAI that could be used also to ease future integration with EOSC services. In this respect eduTeams have been selected by EuroFusion as the proposed solution for community AAI Proxy.

Other partners of the CC have tested the IAM service as well and then integrated it with the IM and PROMINENCE services.

2.3.2.5 Choice of example workflows

The MAST data processing chain consists of over 30 processing codes running as a DAG workflow. In production it has always been executed by a series of scripts on a single machine. It was originally anticipated to run this workflow on different clouds and ensure it could be run within the minimum

[8] https://confluence.egi.eu/display/EOSC/Authentication+and+Authorization+Infrastructure+-+AAI

intershot interval. However, there was a blocking issue due to the extensive use of IDL. As this commercial software requires access to a licence server it is difficult to run on clouds. We attempted to port some of MAST processing codes to the open-source Gnu Data Language (GDL) which is compatible with IDL. While we had some success it turned out not to be possible to get any codes to run to completion due to the use of Dynamically Loadable Modules (DLMs) which are not currently supported in GDL. In future there is expected to be more use of open-source software (Python) rather than IDL but at the moment this is a major blocking issue preventing running these workflows at larger scales on clouds.

Some other fusion workflows at CCFE have a different complication in that they are deeply embedded within custom code and could not easily be extracted to use with a different workflow management system without very significant effort.

A new application for computing a large number of plasma equilibria was developed, in view of the long plasma experiments (lasting several minutes) that will be carried out soon on WEST and later on ITER. It was designed to be run as a map-reduce style workflow. The computation time window is divided into N time intervals, then N independent tasks are computed. The results are merged in one single pulse file when all the tasks are completed. This application leverages both IMAS and Newton direct and Inverse Computation for Equilibrium (NICE), which are routinely used for WEST tokamak operation. Indeed, the WEST intershot processing chain makes systematic use of IMAS and all WEST processed data is stored in the IMAS framework.

A Docker image was developed containing an installation of IMAS, NICE and the OneData client. The initial version of the application consisted of a client, executed by the user who provides various input parameters, which creates all the required processing tasks and uses ssh to execute them in Docker containers on a single remote host. Once the processing jobs complete the client runs a merge job. OneData was used for storing the input data, intermediate files generated by the processing jobs and the final merged output.

2.3.2.6 Running a fusion workflow across multiple clouds

For testing the integration of multi-cloud compute with storage the previously developed scripts were modified and extended to make use of PROMINENCE, enabling the workflow to be much more scalable. In this case the user executes a Python script which generates a file containing the JSON representation of a DAG workflow consisting of a group of processing jobs, almost identical but with different input parameters, followed by a merge job to combine all the outputs. Once the user creates this JSON file they submit it using the PROMINENCE CLI. Again all data access and storage of intermediate files are on OneData. Onezone was provided by EGI DataHub and two OneData providers (in the UK and France) were deployed for the tests. The two providers were joined to the same space, allowing OneData to automatically replicate files, as necessary.

We also made use of OneData for storing the required container images. Two choices for the container runtime were considered, udocker and Singularity, both of which are supported by PROMINENCE. For udocker the image was stored on OneData as a 3.7 GB tarball generated by "docker save", while for Singularity a Singularity Image File (SIF) was created. The choice of container runtime had no effect on the application walltime, but the combined time taken to copy the container image from OneData and create a container was found to be around 5 times faster with Singularity compared to OneData. This was particularly important when using a remote OneData provider, where this took 11 minutes using udocker. So it therefore seems to be preferable to use Singularity rather than udocker for our use case.

There are two ways for an application to read data from OneData: directly read files from the mount provided by Oneclient or to copy the files first to local disk and read from there. For direct reads it turns out to be critical to use a local Oneprovider: for an example processing job reading from a local provider was almost 4 times faster than reading from a remote provider. When data is copied to local disk first the closeness to the provider was less critical - the walltime increased by about 30% when reading from a remote provider compared to local.

We were able to successfully run this WEST workflow using PROMINENCE, distributed across 5 OpenStack clouds in the EGI Federated Cloud, with data access provided by OneData.

During the course of the project use has been made of the following EOSC services:

- B2SAFE
- EGI DataHub
- EGI Federated Cloud
- EGI Check-in
- INDIGO Access and Identity Management (IAM)
- Infrastructure Manager (IM)
- PROMINENCE

For the IAM, IM and PROMINENCE services we used our own deployments rather than services ordered through the EOSC marketplace.

In August 2019, an SLA was signed between EGI providers CESGA and UNIV-LILLE and UKAEA in order to provide compute and storage resources for the Fusion Competence Centre. We also had access to CPU resources at Julich and additional opportunistic EGI Federated Cloud resources through the vo.access.egi.eu VO.

The outcome of this work is a successful proof of concept for WEST intershot processing running on clouds, making use of OneData for data management and PROMINENCE for managing the processing jobs and infrastructure.

The performance gain could be much higher with a more demanding simulation code, as a relatively simple component was used for this proof of concept. It will be beneficial already for present day tokamak machines (e.g. for probabilistic data processing) and even more important for ITER.

Compute resources being close to the data was found to give a huge performance benefit.

More flexibility of OneData replication options would be valuable to avoid replicating the whole tokamak database but to choose selected pulses instead, for example the most recent data.

2.3.3 Lessons learnt

- Users need an X509 certificate in order to join the Fusion VO, but the certificate is not required for any other Fusion activities. This is problematic because users in the Fusion community do not generally have X509 certificates. It turns out that in future it will be possible to convert the Fusion VO into a EGI Check-in based VO which will overcome this difficulty.
- OneData requires exactly the same versions on the client and server and does not throw a helpful error message if the versions are different.

2.3.4 Impact

The CC further developed the PROMINENCE software, deployed it as a service on the access.egi.eu VO, and registered it[9] in the EOSC Portal. The impact just after the proof-of-concept is low at the moment.

Several invited talks have been given about this work, including a talk given to BCS Oxfordshire (The Chartered Institute for IT) in June 2020 and a talk at OpenInfra Day London in April 2019.

2.3.5 Future plans beyond EOSC-hub

The next step is to move the PROMINENCE service into production and use it for CPU-demanding fusion experiment data processing, in particular for WEST.

It is planned to write a paper to be submitted to a journal such as “Fusion Engineering and Design” on the containerised WEST map reduce workflow.

The outcome of the investigation of the AAI systems available in EOSC-hub have been passed to EUROfusion AAI mini project and basing on that the pilot with the selected AAI services (AAI community proxy, IdPs) has been developed. The future deployment of the technology validated in the pilot is currently further discussed in the EUROfusion.

2.4 Marine

2.4.1 Initial ambition (in 2018)

The ocean experts are now converging in the estimation of integrated indicators such as global warming. However these indicators, based on interpolation of unevenly distributed observations, do not describe consistently the climate change. To better understand the ocean circulation and climate machinery, data scientists need to directly access the original observations otherwise diluted in spatial synthesis. Original observations are published by Research Infrastructures (Argo, EMSO, ICOS…) and data aggregators (SeaDataNet, Copernicus Marine…).

The Marine Competence Centre long term ambition is to deploy Ocean observations on EOSC infrastructure for data analytics. The work in the CC focuses on two areas:

[9] PROMINENCE service: https://marketplace.eosc-portal.eu/services/prominence

1. Making Argo data more easily accessible for subsetting and online processing. IFREMER and its partners work in this area.
2. Simplifying/harmonising the access to data that resides at SeaDataNet partners from cloud-based applications. MARIS has worked on this area.

2.5 Euro-Argo

2.5.1 Progress and key results

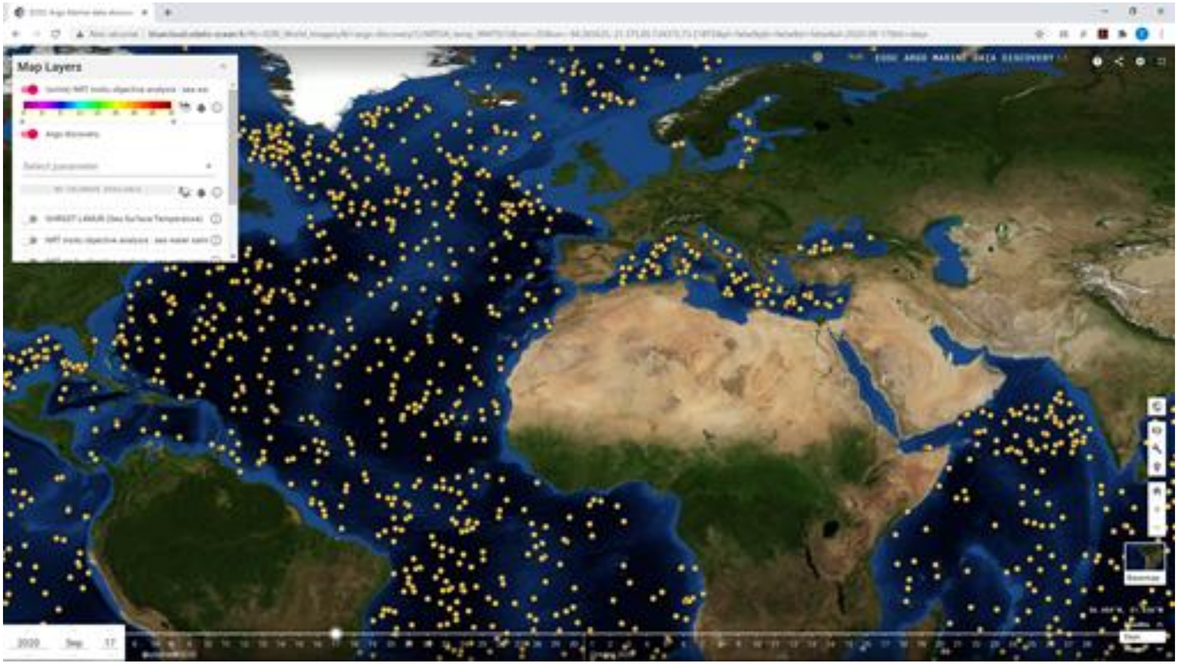
The initial ambition was to set up a workplace on EOSC-hub for ocean experts and data scientists to manage large amounts of data, metadata and information. Within Marine CC, data scientists could work on EOSC-hub with direct access to 20 years of original observations and improve their understanding of global ocean changes. The Marine Competence Centre long term ambition is to push Ocean observations from Research Infrastructure on EOSC infrastructure for data analytics.

This ambition was reasonably addressed. We did set up a Marine Data Analytics Platform with:

- Data storage and computing clusters from EOSC-hub (on Finland EUDAT-CSC and France EGI-LAL-IN2P3 EOSC platforms)
- Jupyter online environment was established for running algorithms codes, save and share notebooks (such as DIVA objective analysis which used Argo and Copernicus CMEMS data).
- A web graphic user interface (GUI) was established for scientists offering the services:
    - Data discovery and visualization of Argo floats data, along with salinity objective analysis, GHRSST sea surface temperature and external satellite data.
- High performance data and metadata access APIs were deployed
    - Metadata API: on top of an Elastisearch index of 10 million records (Argo floats metadata)
    - Data API: on top of a Cassandra NoSQL database of 2 billion records (Argo floats observations)

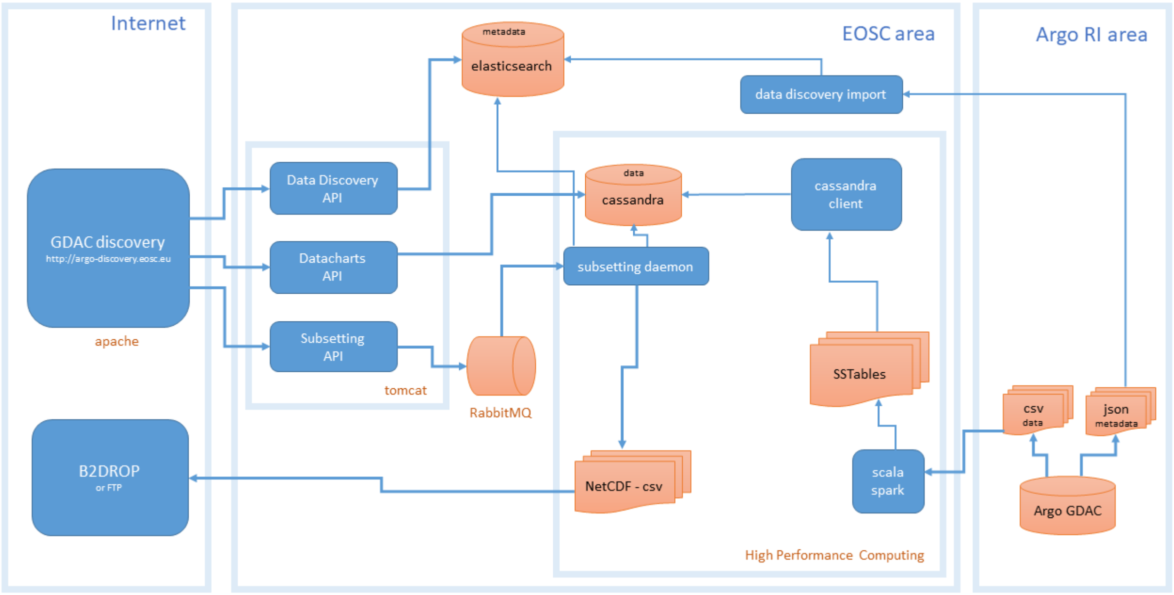
The whole offering was registered in the EOSC Portal-Marketplace as 'Argo floats data discovery service'[10] with the following GUI:

[10] https://marketplace.eosc-portal.eu/services/argo-floats-data-discovery



The Argo data discovery service architecture is based on high performance requests on metadata (10 million records indexed on Elasticsearch) and data (2 billion Argo observations in Cassandra NoSQL database). The data and metadata APIs are publicly available for machine-to-machine applications. The service is hosted in two EOSC-hub cloud provider nodes (CSC in Finland and LAL-IN2P3 in France), the overall architecture is pictured below:

Argo GDAC data discovery and subsetting services



2.5.2 Lessons learnt

- The Docker packaging of complex services such as graphic user interface is an excellent solution for cloud based applications. This technique proved its efficiency and scalability with the deployment of Argo data discovery service on two distinct locations. With Docker simplified packaging and deployments, Marine data service can be pushed closer to the end users infrastructure.
- The use of high-performance query techniques for metadata (Elasticsearch) and data (Cassandra) proved to be robust and scalable. The service validation was performed by CNRS scientists. The response time for requests on 10 million metadata records (Elasticsearch) and requests on 2 billion data records is now instantaneous (e.g. less than one second). On top of these information systems, powerful and efficient API have been developed and included in Argo data discovery service.
- The use of Jupyter Notebooks from EOSC servers using Marine data (Argo, SeaDataNet, Copernicus) to generate scientific products such as DIVA objective analysis proved to be efficient. The Virtual Research Environment for scientists will be based on this combination of shared computing codes (Notebooks) on high performance computing servers (operated from Jupyter) with access to Research Infrastructures data (through APIs or direct dadaste access).

2.5.3 Impact

The ‘Argo floats data discovery service is available fully open access in EOSC and serves data anonymously to users. We are unable to directly track the use of data provided through this in EOSC. However, Euro-Argo yearly bibliographic survey reveals that more than 4500 significant publications acknowledged Argo data, since 2001. In 2020, there was more than one publication a day[11] citing Argo data.

2.5.4 Future plans beyond EOSC-hub

EOSC-hub Marine Competence Centre involved Argo float discovery service in the BlueCloud H2020 project. BlueCloud involves CSC and IN2P3 (the hosting nodes of the service), so the operation is secured there until the end of September 2022.

The service also links to ENVRI-FAIR, where will be used to push Argo data in the cloud for data scientists.

[11] the list of Argo significant publications: https://www.seanoe.org/data/00311/42182/relateddoc.htm, The Argo data bibliographic survey https://archimer.ifremer.fr/doc/00511/62234/

2.6 SeaDataNet

2.6.1 Progress and key results

The SeaDataNet network consists of the main European national oceanographic data centers making their data archives available in a harmonised way. Therefore the main topics of the SeaDataNet community are:

1. Downstream: Providing a cloud platform with common services for data pre-processing, a Virtual Research Environment (https://www.seadatanet.org/Software/VRE), subsetting, analyses, visualizations, publishing, DOIs.
2. Standards as glue: Applying common standards and interoperability solutions for providing harmonised data and metadata.
3. Upstream: Providing harmonised discovery and access to data output from multiple sources, such as European research and monitoring data gathering, but also from other European and international data infrastructures

The focus of the SeaDataNet work in the marine CC has been on access to the data and sharing, in line with the initial ambition with an additional exploration in performance increase in data viewing of discovered data.

The basis for the upstream services on SeaDataNet is the so-called CDI system, one of the core services of the SeaDataNet infrastructure. It provides a highly detailed insight and unified access to the large volumes of marine and oceanographic data sets managed by the distributed data centres. It is a fine-grained index (ISO 19115 – ISO 19139) to individual data measurements (such as a CTD cast or moored instrument record).

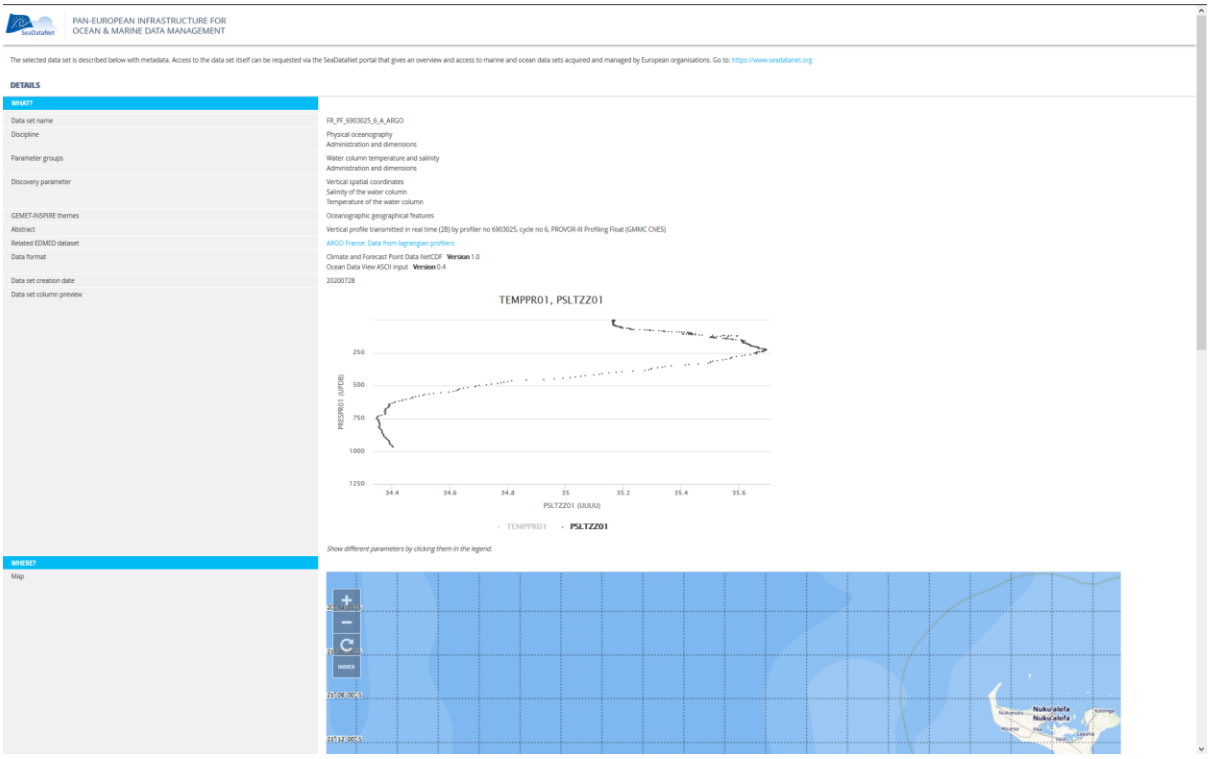
Current content is 2.5M metadata records which are describing and are linked to observation datasets. Around 89% of these files have a light restriction and are stored and available directly from the EUDAT cloud. The datasets come from more than 110 data centers (providing data from over 730 data originators) connected using Replication Manager to publish data.

The current challenges of this system are:

- Managing the access to the distributed data published via the 110 nodes that are almost 90% stored in a cloud. Providing access to this “data lake” of metadata and datasets from the cloud toward (internal), e.g. the SDN Virtual Research Environment, and (external) to applications from other communities and EOSC .
- Performance of visualisation of datasets and dataset collections directly from the cloud because the files are small but many, and have many values (e.g. a time-series or depth profile)

In order to investigate if there are EOSC-hub components available that can overcome the challenges the work has focussed on an alternative way to share and provide access to the data by publishing the datafiles on OneData nodes. 10K files from 5 different data centers have been loaded on two different OneData nodes to explore the concept: at INCD/LIP in Portugal, and IFCA in Spain. After a short learning curve to work with OneData this worked smoothly, and it was eventually possible to get from a central location a virtual overview of data on the distributed nodes as if they were centrally available. It has also been explored how the synchronisation of files from a datanode could work and how metadata can be added to control versioning.

As a second step a Cassandra database has been configured that can handle, store and query the incoming datasets for visualisation. This work was done in close contact with EOSC-hub Marine CC partner IFREMER that had prior experience with this component as part of ARGO work. Scripts were created to load the datafile values directly from the OneData nodes into the Cassandra database. And to complete the concept the Cassandra database was queried to view the data directly in a test interface with convincing very high performance (almost immediate response). See the below screenshot of the result:



PAN-EUROPEAN INFRASTRUCTURE FOR
OCEAN & MARINE DATA MANAGEMENT

The selected data set is described below with metadata. Access to the data set itself can be requested via the SeaDataNet portal that gives an overview and access to marine and ocean data sets acquired and managed by European organisations. Go to: https://www.seadatanet.org

DETAILS

WHAT?	
Data set name	FR_PF_6903025_6_A_ARGO
Discipline	Physical oceanography Administration and dimensions
Parameter groups	Water column temperature and salinity Administration and dimensions
Discovery parameter	Vertical spatial coordinates Salinity of the water column Temperature of the water column
GEMET-INSPIRE themes	Oceanographic geographical features
Abstract	Vertical profile transmitted in real time (2B) by profiler no 6903025, cycle no 6, PROVOR-III Profiling Float (GMMC CNES)
Related EDMED dataset	ARGO France: Data from lagrangian profilers
Data format	Climate and Forecast Point Data NetCDF **Version** 1.0 Ocean Data View ASCII input **Version** 0.4
Data set creation date	20200728
Data set column preview	

Show different parameters by clicking them in the legend.

WHERE?

Map

2.6.2 Lessons learnt

The OneData system with distributed nodes functions very well, offering live and accurate overview of available files on nodes and transfer to cloud. More research is required for handling-controlled access to potentially providing access to collections of restricted datasets (that should only be released after negotiation for certain users but kept secure in the meantime). OneData could be a very good way to share SeaDataNet collections to external systems, but also as a kind of data lake to be used internally for the SDC VRE, where the VRE could then also connect to OneData data lakes of other data infrastructures.

The access to the datafiles when loading to the database was somewhat slow, but it must be mentioned that it was not optimised yet. This may be increased by e.g. using parallel actions.

Regarding the performance of visualisation, the Cassandra database querying is fast but not very flexible. The database structure needs to be decided in advance, meaning that later changes in the way the user might want to query and view the data might lead to a big change in the system. Alternatives with other DBs and Elastic might fit as well and should be explored. Visualising the data from Cassandra when set up well to the query, was very responsive.

2.6.3 Impact

Since this work is really in pilot phase, and OneData was already a registered service in EOSC, no additional services have been registered by the SeaDataNet community of the CC. However, the value of the action has been to test the services as a potential component for the publishing and accessing data from the SeaDataNet domain. The results have been presented by Peter Thijsse (MARIS) during the EGI conference, and an abstract has been submitted to the SeaDataNet IMDIS conference in April 2021, where usually an audience of 200-300 persons from the marine data management domain will be present.

2.6.4 Future plans beyond EOSC-hub

More research and testing will be required for possibly handling controlled access to a data lake with restricted datasets. And it would be interesting to explore the option to use the OneData nodes as input for the SDC VRE where data can be processed and combined with other sources, e.g. using a Jupyter notebook. But first results are positive, OneData could be a very good way to share datasets and collections and should be explored further.

As the next step a prototype could be explored in a smaller research project, e.g. involving the SDC VRE working group. Or it could be explored at a larger scale in a future SeaDataNet project to share the data products and collections in this way for use in EGI/EOSC related central VRE services.

In Jan 2021 MARIS will be involved in the EGI-ACE project (Advanced Computing for EOSC) and will deploy the SeaDataNet WebODV application at the EGI cloud infrastructure and will mobilise its large scientific user basis to adopt the WebODV cloud version for online analytics. The use of OneData in the setup will be also assessed.

2.8 EISCAT_3D

2.8.1 Initial ambition (in 2018)

EISCAT Scientific Association participates in the EOSC-hub WP8 Competence Centre with the aim of developing a data portal for users of the future radar system EISCAT_3D, which is planned to start operation during 2021-2022. The aim of the CC is to have a working prototype open for public access during EOSC-hub. In order for this to be achieved necessary tools, services and infrastructures need to be deployed and integrated for functional data management and data processing. DIRAC interware will be integrated and used as a portal and function as a single access point towards e-infrastructures. The portal should authenticate and authorize the future users of EISCAT_3D data and allow them to discover data and submit data sets for analysis. EUDAT B2 services are intended to be integrated and unify EISCAT_3D data discovery and access across possibly different redundant storages. EGI and INDIGO services will be integrated, deploying the software stack on HPC/HTC systems including release management.

2.8.2 Progress and key results

2.8.2.1 DIRAC storage and job management

The EOSC-hub CC for EISCAT_3D has built on the outcomes of the EGI-Engage competence centre and thus concentrated around the DIRAC Interware system. The inherited setup consisted of

- DIRAC Storage Element (SE)
- DIRAC File Catalogue
- DIRAC Web GUI
- Compute resources

Access to the DIRAC Web GUI has been through X.509 authentication. This has turned out to be a major hurdle for normal users, so we are changing to federated ID. This is especially important for EISCAT's non-European users including members in China and Japan. So during EOSC-hub the CC integrated Check-in into DIRAC for user authentication and adopted Perun for user attribute management and authorisation. There will be at least one group manager per EISCAT member, with rights to approve group membership.

Job submissions are at the present stage deployed on the cPouta cloud at CSC (a CC member) by starting docker containers on the assigned OpenStack VM(s). Existing EISCAT user software is integrated: the general RTG tool (Octave, C) for displaying Level 2 data (but not L3). Lag profiling (recreating L2 data from those experiments that sample L1 data) is also in progress. The general analysis software GUISDAP (creating L3 data from L1/L2) has not been integrated.

2.8.2.2 AAI

Data embargo rules and system operation requirements demand a fine-grained control of user authorization. There are also users from around the world including China and Japan, so we have to accommodate for several identity provider options including social logins.

EGI Checkin was selected as SAML2.0 / OIDC identity proxy and has been tested with institutional and social logins. User authorization will utilize group attributes provided by EGI Perun.

2.8.2.3 EGI resources

During the project EISCAT has made use of the following services provided by EOSC and partners:

- DIRAC File Catalogue (SQL database) and Web GUI are hosted at Cyfronet, Poland
- EGI Check-in is used as ID broker
- perun.egi.eu is used for user attribute management

During the prototyping, the DIRAC Storage Element has been running in EISCAT's own facilities and provided access to data from the legacy EISCAT radars, which are the existing systems that will be replaced by EISCAT_3D in near future. We have also used cloud compute resources provided by CSC.

2.8.2.4 EOSC Marketplace registration

The EISCAT portal was released as an EOSC thematic service and is listed on the EOSC Marketplace[12]. EISCAT is running a series of Access mini workshops (online) with selected future users of EISCAT_3D. This will be further reported in WP11. The procedure of registration, login and use has been tested and feedback provided to the developers of DIRAC, Check-in and Perun.

2.8.2.5 Overview of prototype system

Figure 2 shows the components of the EISCAT Data Portal Prototype and also indicates the planned metadata harvesting to B2FIND and use of the Check-in OAuth service for accessing Jupyter notebooks (see Future plans below).

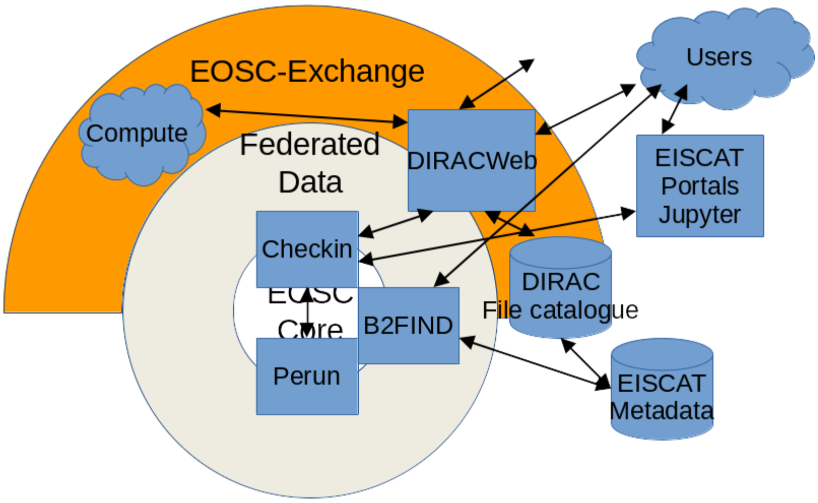


Fig.2 – Components of the EISCAT Data Portal Prototype

[12] EISCAT Data Access Portal: https://marketplace.eosc-portal.eu/services/eiscat-data-access-portal

2.8.3 Lessons learnt

This CC has allowed us to gain much experience in fields which are generally not familiar to scientists. In summary

- Bugs, configuration problems, and questions regarding use of services are being encountered continually. A clear procedure for feedback is essential. Ticketing systems have proven to be valuable.
- X.509 authentication and where to obtain personal certificates is likewise unknown to typical users. The integration of federated identity into the portal was essential.
- Metadata for user interaction (data discovery, FAIRness) is different from the metadata required for the internal operation of a system (from the system monitoring level and up to location and sorting of raw data objects) so different approaches to metadata will be necessary.

2.8.4 Impact

The prototype service is registered in the EOSC portal and available through the Marketplace. During the testing of the prototype we have reached out to users from the Swedish Institute of Space Physics and the University of Oulu, Finland, which are both major users of EISCAT data. The feedback has allowed improvements and modifications of Check-in, Perun and DiracWeb.

2.8.4.1 Outreach and training

Regular participants in our data access prototyping workshops are from:

- Swedish Institute of Space Physics (IRF), Kiruna department, 2-3 persons
- IRF Uppsala department, 1 person (SE data contact / group manager)
- University of Oulu Physics department, 2 persons (1 FI data contact / group manager)
- University of Oulu Sodankylä Geophysical Observatory 1 person
- GRNET (Check-in developers) 1-2 persons
- CESNET (Perun developers) 1 person
- EGI 1-3 persons

2.8.5 Future plans beyond EOSC-hub

In close connection with this CC work we have initiated these collaborations

- Metadata harvesting on experiment collection level. Existing EISCAT and future EISCAT_3D metadata to be ingested into B2FIND and other services. Project: mainly ENVRI-Fair.
- Use Check-in as an identity broker for other EISCAT services including Jupyter Projects. The source of funding for this work is still to be determined.
- Use storage, data transfer and compute provided by Nordic e-infra providers. Close collaboration with NeIC and LHC NT1. Funding through EISCAT_3D project WP3.
- Data replication to members will be enabled. The source of funding for this will be on a national level, and work for it is underway in Japan, Norway and the UK.

Two EOSC-related projects with EISCAT involvement have been recently accepted: EGI-ACE and PITHIA-NRF. We yet to understand how these can bring the development of the EISCAT Portal further.

2.9 EPOS-ORFEUS

2.9.1 Initial ambition (in 2018)

The CC drives collaboration between EOSC-hub and the ORFEUS-EIDA federation of EPOS. The CC collects and assesses the requirements of the solid-Earth science community, with a specific focus on Seismology, and addresses them by leveraging the EOSC-hub technical offerings. The CC delivers a software platform that facilitates access and exploitation of computational resources; it supports and fosters harmonisation of best practices for data management at ORFEUS-EIDA; and it enables the generation of seismological products customised on user requirements. By the end of the EOSC-hub project the CC aims to have a pre-production quality, modular software platform that could be deployed at (selected) data centres. However, the actual deployments will depend on agreements for service provisioning and operation.

2.9.2 Progress and key results

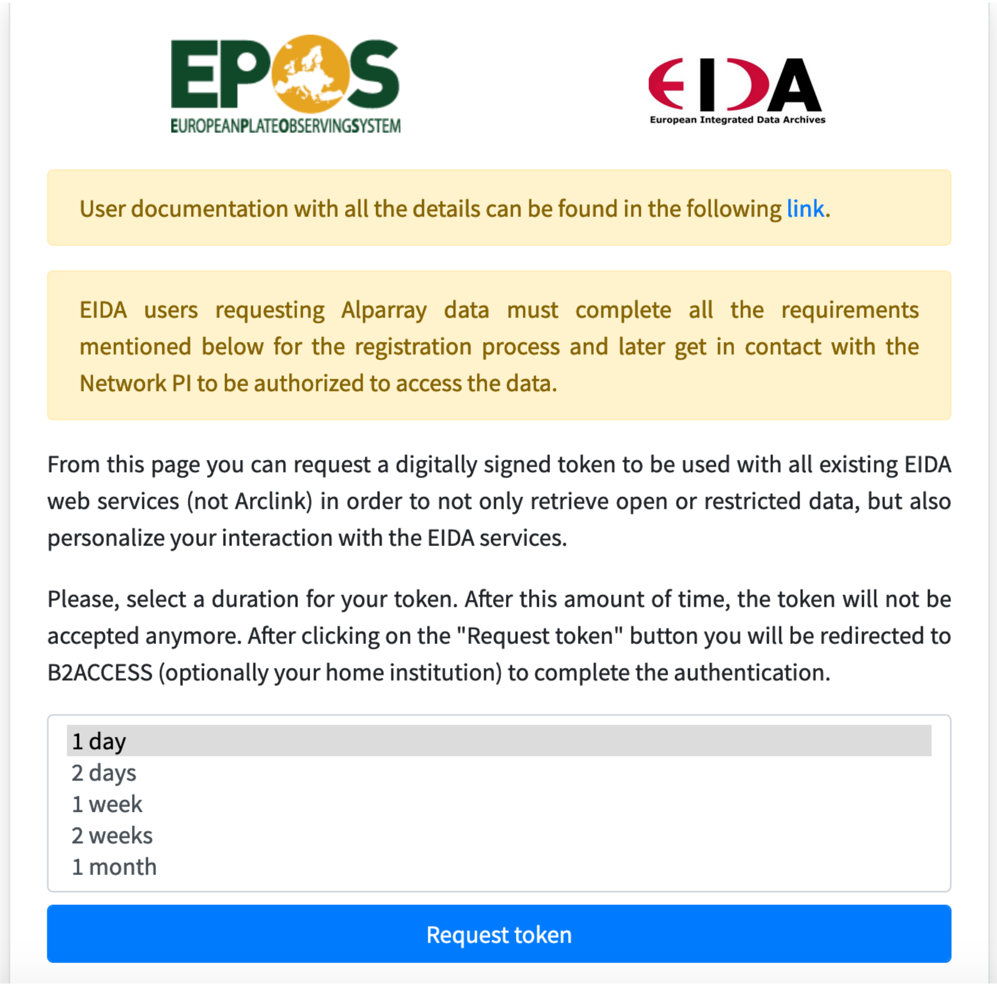
The activities carried out in the EPOS-ORFEUS-CC focused on four areas. The activities are described in the next subsections.

2.9.2.1 AAI integration

A first goal of the CC addressed a primary requirement identified by the EPOS Seismology community: an operational, scalable, federated AAI service that is able to interoperate with community services and is compliant with state-of-the-art technologies. Such a goal was achieved by devising a solution (EAS[13]) based on B2ACCESS. The developed solution was tested in a focused use case by a targeted seismological community (AlpArray use case). Successively, the service was rolled out in production and it is currently operated by ORFEUS-EIDA. This was a major achievement of the project.

Despite that almost all the data hosted by the data centers are open and accessible without the need of being authenticated, more than 400 users have already adopted the system from a base of around 2500 global users. Part of the effort was also used to enhance third-party client tools, broadly used within the community, in order that those tools can support our new AuthN/AuthZ system (e.g. fdsnws_fetch, Obspy). Today, seismologists can access waveform data without the need to make a distinction between open and restricted with the same tools they have been working with, just making use of the tokens provided by EAS. Figure3 shows the web interface of the token requesting system.

[13] https://geofon.gfz-potsdam.de/eas/



EPOS
EUROPEANPLATEOBSERVINGSYSTEM
EIDA
European Integrated Data Archives

User documentation with all the details can be found in the following link.

EIDA users requesting Alparray data must complete all the requirements mentioned below for the registration process and later get in contact with the Network PI to be authorized to access the data.

From this page you can request a digitally signed token to be used with all existing EIDA web services (not Arclink) in order to not only retrieve open or restricted data, but also personalize your interaction with the EIDA services.

Please, select a duration for your token. After this amount of time, the token will not be accepted anymore. After clicking on the "Request token" button you will be redirected to B2ACCESS (optionally your home institution) to complete the authentication.

Fig. 3 – web interface of the token requesting system

2.9.2.2 Data staging

A second goal focused on the research of technological solutions to facilitate analysis and computation of seismological data on the cloud. This encompasses two aspects:

1) Enabling staging of data from institutional seismological archives to external compute resources.

   AND

2) Evaluating and selecting usable compute facilities that fulfil the requirements of typical use cases.

With the respect to 1) the CC investigated different solutions by adopting both community services and standard APIs (e.g. WFCatalog and FDSN WS) and data services from the EOSC portfolio: B2HANDLE, B2SAFE and B2STAGE (e.g. airodsAPI[14]).

[14] https://github.com/massimo1962/http-api-airods

Preliminary activities were performed to enable the data staging use case. Those include the assignment of PIDs to seismic waveform data and the replication of seismological data archives to EOSC-hub providers' facilities.

A typical data management workflow has been implemented (with slight variations) in the participating seismological EIDA data centres, entailing the following steps (Fig.4):

1. Seismic waveform data (daily MSEED files) registration to local iRODS service within B2SAFE.
2. PID creation for the registered data using the B2HANDLE service.
3. Data replication to EOSC-hub providers' facilities using the B2SAFE service.
4. PID creation for the replicated data using the B2HANDLE service.

Moreover, a community catalogue i.e.: WFCatalog has been tested in order to associate the minted PIDs to community metadata.

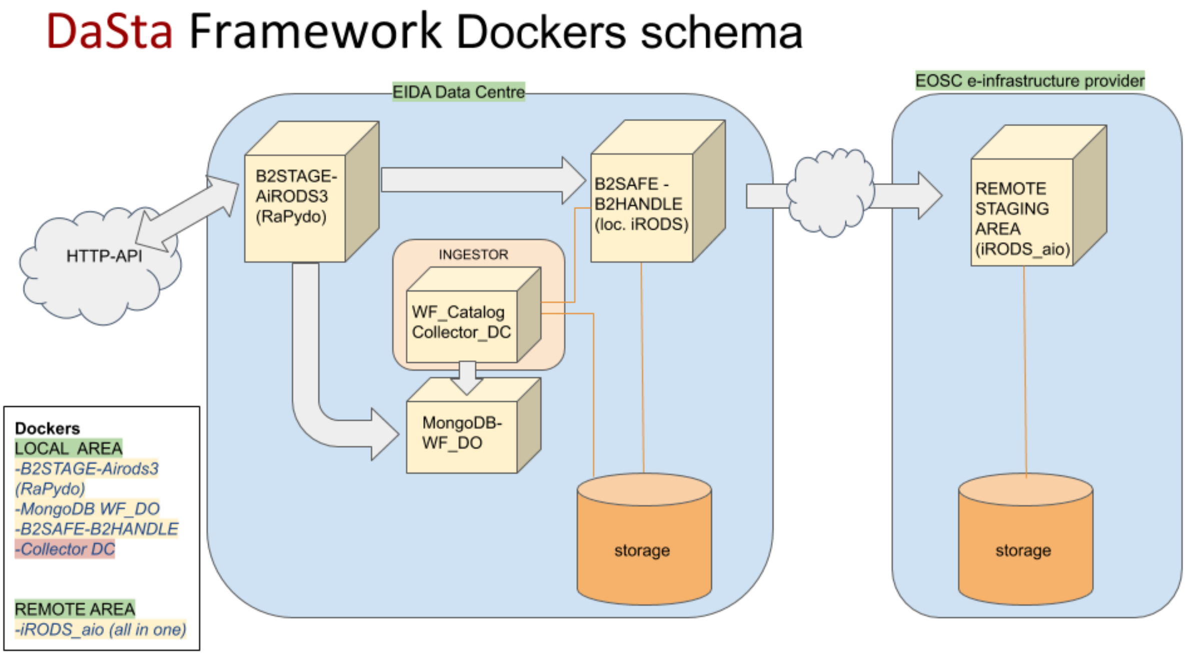


Fig. 4 – DaSta Framework Dockers Schema

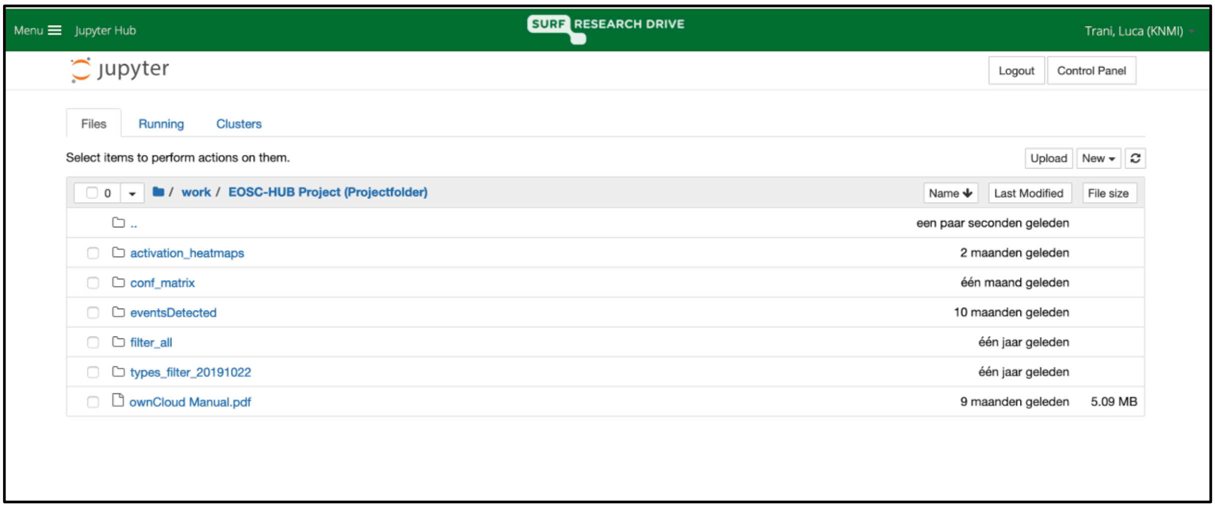
The pilot yielded valuable results and developed useful knowledge to advance the service offering of the EPOS Seismology community. The CC analysis concluded that:

- Community standard APIs based on the HTTP protocol remain the preferred way to access data as they offer a number of advantages: they are widespread in the seismological community (reduced buy-in costs), well integrated in community tools (e.g. Obspy), their application paradigms are known (ease of operation and maintenance). However, they might suffer from inherent technical issues when scaling up to higher volumes of data/requests.

- The alternative technologies can help solve the scaling problems; however, they need to overcome some barriers for adoption in order to make them available at all data centres and to cover additional operational and maintenance costs. Also, some issues remain to be solved at the level of the EOSC Core service e.g. by better supporting the integration of distributed data services and by facilitating integration of providers' compute nodes. (See in next section)

2.9.2.3 Evaluating compute facilities

Concerning part 2) of our second goal different computing environments were investigated. They include solutions based on the Jupyter technology (KIT Jupyter, NOA-GRNET Jupyter, SURF ResearchDrive); and HPC Cloud and Custom Cloud Solutions (CCS) by SURF.



Select items to perform actions on them.

□ 0 / work / EOSC-HUB Project (Projectfolder)	Name	Last Modified	File size
..		een paar seconden geleden	
□ activation_heatmaps		2 maanden geleden	
□ conf_matrix		één maand geleden	
□ eventsDetected		10 maanden geleden	
□ filter_all		één jaar geleden	
□ types_filter_20191022		één jaar geleden	
□ ownCloud Manual.pdf		9 maanden geleden	5.09 MB

Integrations with external services such as EGI Check-in (AAI), B2DROP and B2SAFE (storage) were also tested. The assessment highlighted different degrees of readiness and capabilities to fulfil the requirements of the EPOS Seismology community: Jupyter based solution proved useful with applications of focused scope and limited capacity requirements. In particular, they are fit for purpose when applied to applications with iterative analysis of the same dataset, which are rather small seismic processing and analysis tasks.

- The setup at KIT included 64GB of RAM, up to 40 CPU cores, 8 GPU cores. As Jupyter notebooks run on their HPC-cluster facility the resources can be also easily extended.
- The tested solution based on Jupyter integrated in ResearchDrive running at SURFsara was deployed on a virtual server with 16 CPU cores and 90GB RAM.
- Jupyter is also available at NOA-GRNET in a Dockerised environment with RAM of 16 GB and storage of 30 GB.

An investigation was carried out to run a data science application based on Machine Learning and Convolutional Neural Network for seismo-acoustic event classification (Trani et al, 2020[15]). That use

[15] Trani L., Pagani G. A., Perreira Zanetti J. P., Chapeland C. G. M., Evers L. "Deep-Quake An application of CNN for seismo-acoustic event classification in The Netherlands", submitted to GRL.

case required a different technological solution supported by a larger infrastructure (i.e. HPC Cloud), because the analysis requires more complex data management with multiple, and ad-hoc access to data at various sites. Finally, in the last part of the project an investigation has been initiated to deploy a provenance-aware API developed and adopted by the EPOS community (SWIRRL[16]) on EOSC compute facilities.

The integration of distributed Jupyter computing environments was indicated as one of the initial goals but eventually it was considered as too challenging and not feasible with the current EOSC architecture. (E.g. B2DROP integration was supported in some Jupyters, but not all of them. AAI was not exactly the same in all of them. And none of the Jypyter providers was willing to adopt the other site's setup because they serve multiple communities.) Such a scenario should be considered and implemented in future developments of the EOSC Core services. Also, the EOSC Core would be the right place where to solve integration and compatibility issues that in this phase were left to the community.

We can conclude that EOSC would have the capability to fulfil the requirements of the EPOS-ORFEUS community by offering a variety of viable technological solutions. However, important technical issues need to be resolved at the level of the EOSC Core service e.g. harmonisation and interoperability of services of the EOSC portfolio. Also, some concerns remain regarding the service provisioning model and the sustainability of the infrastructure.

2.9.2.4 Rule-based data management

Finally a data management framework (RuleManager[17]) was developed to implement configurable policies at seismological data centres (See concept Fig 5). RuleManager is able to interoperate and integrate with community services (e.g. WFCatalog) and EOSC services (B2HANDLE, B2SAFE).

Data management policies to implement PID registration and archive replication have been piloted and demonstrated the feasibility of RuleManager to address the requirements of the EPOS Seismology community.

[16] https://gitlab.com/KNMI-OSS/swirrl/swirrl-api → EOSC version https://eosc.swirrl.k8s.surfsara.nl/swirrl-api/#

[17] https://github.com/KNMI/DMPilot-RuleManager

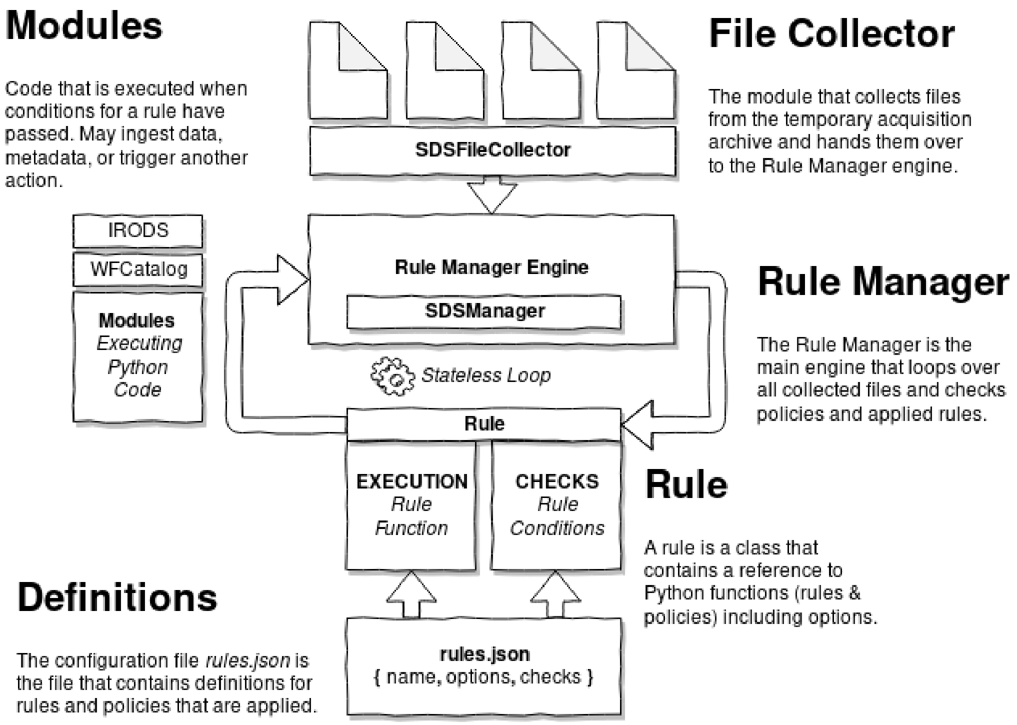

Modules
Code that is executed when conditions for a rule have passed. May ingest data, metadata, or trigger another action.
IRODS
WFCatalog
Modules
Executing
Python
Code
SDSFileCollector
File Collector
The module that collects files from the temporary acquisition archive and hands them over to the Rule Manager engine.
Rule Manager Engine
SDSManager
Stateless Loop
Rule Manager
The Rule Manager is the main engine that loops over all collected files and checks policies and applied rules.
Rule
EXECUTION
Rule
Function
CHECKS
Rule
Conditions
Rule
A rule is a class that contains a reference to Python functions (rules & policies) including options.
Definitions
The configuration file rules.json is the file that contains definitions for rules and policies that are applied.
rules.json
{ name, options, checks }

Fig. 5 - RuleManager

2.9.3 Lessons learnt

- The CC reached an operational, scalable, federated AAI service that is able to interoperate with community services and is compliant with state-of-the-art technologies. This is a major achievement of the project.
- The Jupyter providers we worked with, although provide similar features, do not look/behave exactly the same, therefore the setup of a system, with multiple Jupyter endpoints each providing one entry point for users from the region, failed. What role will EOSC play in harmonising the setups? Can the EOSC interoperability guidelines help? (They will not solve differences in feature sets, but can serve as a baseline across communities. EOSC should provide integration support on top of this baseline.)

2.9.4 Impact

The EPOS-ORFEUS-CC delivered important results in terms of knowledge building and sharing, service development and infrastructure enhancement. The investigations performed and the achievements of the CC enabled the EPOS Seismology community to pilot and pioneer the integration with EOSC and showed concrete paths for future collaborations. Critical issues were also highlighted, those will require further research and additional activities.

The results of the CC were made available to the community via 11 outreach and dissemination events[18] in the form of 1 poster 8 presentations, 2 webinar talks. The target audience was chosen depending on the service maturity level.

The CC contributed to establish a crucial operational service such as the federated AAI system: EAS. The service was initially tested on a selected user group and then made public and extended to the whole ORFEUS-EIDA community. As mentioned previously, the service has been for a long time in production and more than 400 users have already adopted it. Moreover, it is also supported by the client tools most used within our community, what was a key factor for their successful adoption. User documentation[19] is available from the web page of the service.

EIDA data centres operators were targeted to transfer knowledge about replication of seismological archives and assignment of PIDs to seismic waveforms. Eventually an additional EIDA node (NOA) implemented that service and others showed their interests. Documentation has been produced internally for the data centres to facilitate the uptake in other EIDA nodes.

The data management policy framework Rule Manager was presented to the EIDA data archive operators and tested in three EIDA nodes (KNMI, NOA and INGV). The code has been made available on a public repository[20] and its development will be continued beyond the project.

Data access and staging methods have been tested and demonstrated in different data centres and presented in workshops and thematic webinars. In the last part of 2020, two webinars were offered to the community. The first one (Nov 24th) on Data Access had a participation of ca. 140 attendants from all around the world. The second one (Dec 14th) focussed on the designed AAI solution is expected to have a similar attendance. Resources related to these webinars can be found in the ORFEUS website[21].

Similarly, presentations were organised to show the capabilities of the Jupyter based solutions for data analysis and processing. Additional webinars on the topic are planned for the first quarter of 2021. In some cases (e.g. GFZ-KIT) those services were made available to selected users and use cases.

Finally, the achievements of the CC include contributions to international conferences and publications[22].

2.9.5 Future plans beyond EOSC-hub

The experience of the EPOS-ORFEUS-CC demonstrated that the established collaboration is beneficial both for EPOS RI and EOSC service providers. It showed how joint efforts can be combined

[18] The events include EPOS Seismology Workshop and ORFEUS Annual Meeting, EGU, EPOS Seismology Meeting 2019, 2019 EPOS Seismology Workshop and ORFEUS Annual Meeting. The full list is available at https://wiki.eosc-hub.eu/display/EOSC/Dissemination+Activities

[19] https://geofon.gfz-potsdam.de/eas/EIDAAuthenticationService.pdf

[20] https://github.com/KNMI/DMPilot-RuleManager

[21] http://orfeus-eu.org/other/workshops/

[22] https://wiki.eosc-hub.eu/display/EOSC/Dissemination+Activities

in order to offer an enhanced service provisioning shaped on user requirements and to tackle new data challenges.

The AAI system will be maintained in production by GFZ, serving the EPOS community and beyond. Webinars/meetings are planned with the international federation of seismology to use this as an international standard (e.g. FDSN standard) for data access.

The EPOS-ORFEUS community has no plans to go forward towards a harmonised European landscape with homogenised Jupyter installations. However the KIT Jupyter system will be finalised with the involvement of early adopter researchers in 2021. The NOA-GRNET data replication setup will be maintained in production by NOA.

From a technical point of view a number of activities will be continued in the context of the EPOS ERIC and EU projects (e.g. ENVRI-FAIR):

- The establishment of the FAIR principles suggests that the development of PID-based services will be extended to additional seismological data centres to form a solid backbone for advanced FAIR data services.
- The development of Rule Manager will be continued and rolled out as an operational framework.
- The solutions for Jupyter-based data analysis and their integration with SWIRRL will be further developed as part of the implementation of the EPOS ICS-D and ENVRI-FAIR.

EPOS participates in the EOSC Future project from mid-2021. The collaboration with e-infrastructures is expected to continue there - although maybe on different use cases than those from the CC.

2.11 Radioastronomy

2.11.1 Initial ambition (in 2018)

The Radio Astronomy CC (RACC) aims to support researchers to find, access, manage, and process data produced by the International LOFAR Telescope. It aims to lower the technology threshold for the Radio Astronomical community in exploiting resources and services provided by the EOSC. Particular aspects that will be addressed are federated single sign-on access to services in a distributed environment and support for data-intensive processing workflows on EOSC infrastructure, notably having access to user workspace connected to high-throughput processing systems, offer portable application deployment, and provide integrated access to a FAIR science data repository. The community is to be empowered to optimally profit from these and increase the science output from multi-petabyte radio astronomical data archives of current and future instruments. The RACC will achieve this by undertaking activities including integration with available federation and data discovery services.

Users will be provided with access to large-scale workspace storage facilities within the EOSC-hub to store and share temporary data and products from pipelines. RACC will empower science groups to deploy their own processing workflows. The RACC lessons learned will serve as input for the design and construction of a European Science Data Center for the Square Kilometre Array (SKA), e.g. via the complementary AENEAS and ESCAPE projects.

2.11.2 Progress and key results

In line with the original ambitions, the RACC has focused on three areas of providing open data services for research based on radio astronomical data:

1. Community access through Federated AAI
2. Portable data processing workflows for LOFAR data processing
3. Improving FAIR data access to high level science data products

Internal resourcing challenges did result in a late start of the activities, but then good progress has been made over the course of 2019 and 2020 and the first results at prototyping and demonstration level have been achieved in all of these areas. The achievements, including the main findings, are detailed in the following sections.

2.11.2.1 Community access through Federated AAI

As part of the EOSCpilot project, a federated AAI environment has been set up for the LOFAR community based on the EGI Check-in service using CoManage. A few customizations had been realized for the Virtual Organization for LOFAR/ASTRON, notably a self-registration endpoint for users to apply for membership of the LOFAR/ASTRON virtual organization, and an enrolment page for members of the community to request group membership in specific roles and/or scientific collaboration groups.

At the start of the EOSC-hub 8.6 activity (the RACC) the following integrations were achieved:

- integration of web-based services such as a Django REST web server using OIDC

- the integration of the JIRA-based ASTRON ticketing system using SAML
- the provisioning of an LDAP server with public SSH keys to enable user access to a Linux portal and SVN code repositories.

After that phase was finished several critical existing user services were still to be integrated with Check-in. These systems required deeper investigations due to their dependency on legacy platforms that do not support SAML, OIDC, or SSH key based access. The legacy services support the (1) submission of science proposals, (2) requesting telescope and/or data archive resources, (3) the administration for managing allocated science projects as they are executed, and (4) the access to the LOFAR Long Term Archive (LTA). The CC continued the Check-in integration with three of these critical service platforms:

- the proposal submission application (1)
- the project management application (3)
- the LOFAR Long Term Archive (LTA) (4)

The proposal submission and project management applications are JAVA applications running in a Tomcat web server. To enable OIDC-based integration with EGI Check-in, the applications were upgraded to a recent version of the Tomcat web container, following which integration with EGI Check-in could be realized by configuring the tomcat-oidcauth module.

For LTA access, a web-based portal is in use that is built on the Python-based AstroWise environment backed by an Oracle database. It was possible to realize OIDC-based integration for this service through application of the oidcrp python module.

In support of the RACC, SURFsara has demonstrated the integration of their Spider compute cluster with federated AAI through provisioning of an LDAP server from the federated AAI provider.

See the overall architecture of the services that are integrated with federated AAI in Figure 6. The architecture shows two types of federations: The left side is with the use of Check-in, the right side is with the use of SRAM. (See details described later).

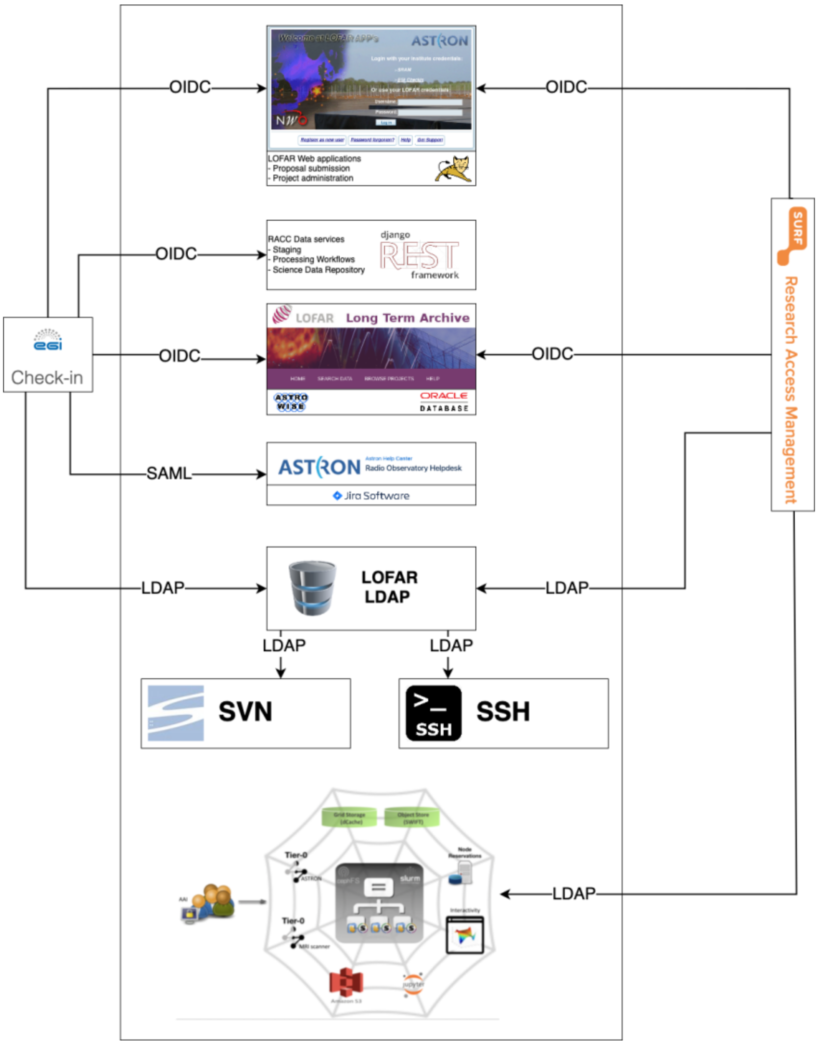
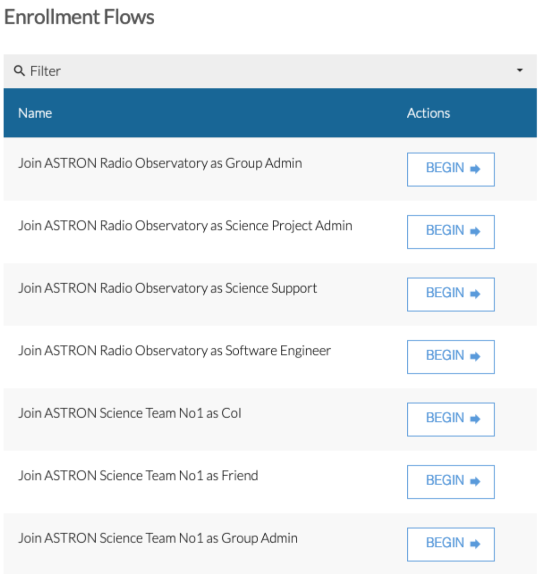


Fig. 6 - overall architecture of the services that are integrated with federated AAI

Two other existing LOFAR data services provide staging functionality, allowing users to request archived files to be retrieved from tape such that they can be copied to an external system. It has been decided to develop a replacement service for the stager based on the python Django REST framework supporting OIDC out of the box. Data retrieval from the LTA is currently supported based on the gridFTP (based on X509 personal certificates) or through a streaming web server hosted at each LTA site where access is provided through integration with the central LOFAR LDAP server.

For data retrieval, the RACC decided to move towards WEBDAV access which is now offered by the dCache system itself and where access can be provided through tokens. The tokens can be generated by the new stager service, so no further effort is required to integrate the data retrieval service with federated AAI.

The EGI Check-in set-up has been evaluated by ASTRON system administrators, software developers, and demonstrated at an internal workshop including representatives from the ASTRON astronomy group. It fulfils the main requirements to support LOFAR community services, but criticism was received about the user-friendliness of the CoManage web interface and the complexity of the enrollment process. (See Fig.7 for a screenshot about the user page.) In its current state for LOFAR, the enrollment interface does not scale to the level that will be required for production deployment as users are offered all available combinations of groups, roles and collaborations.



Enrollment Flows

Filter

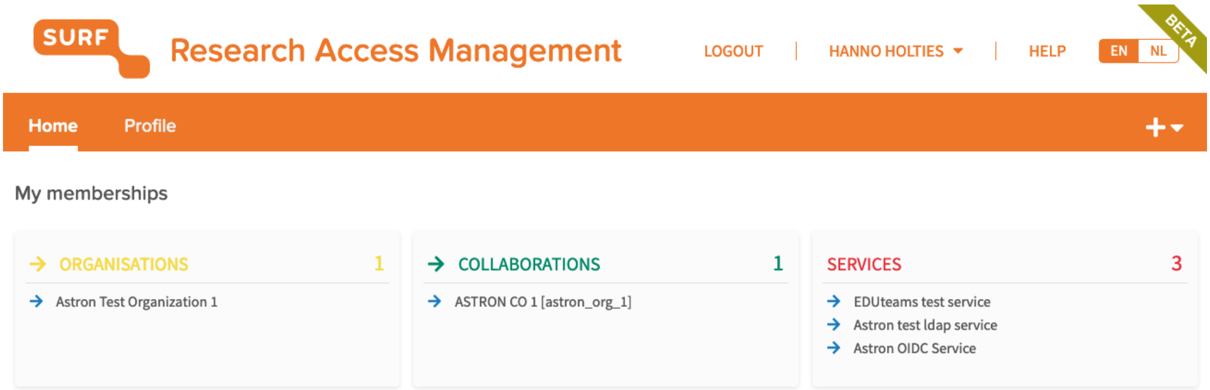
Name	Actions
Join ASTRON Radio Observatory as Group Admin	BEGIN ➔
Join ASTRON Radio Observatory as Science Project Admin	BEGIN ➔
Join ASTRON Radio Observatory as Science Support	BEGIN ➔
Join ASTRON Radio Observatory as Software Engineer	BEGIN ➔
Join ASTRON Science Team No1 as CoI	BEGIN ➔
Join ASTRON Science Team No1 as Friend	BEGIN ➔
Join ASTRON Science Team No1 as Group Admin	BEGIN ➔

Fig. 7 – enrollement interface User page

On the positive side, the CoManage environment of EGI Check-in allows for extensive detailed configuration, and support for LOFAR by GRNET has been very good. Also, the OIDC self-service allows integrating new services in an agile manner without the need for action to be taken by central EGI Check-in administrators. This saves valuable time in particular during the development/evaluation phase. Finally, CoManage also supports LDAP provisioning configurable by a community administrator. All in all, Check-in (CoManage) is highly configurable and allows for a high level of autonomous community management.

Although further work would be possible to improve the usability and scalability of the web interfaces, the technology-oriented design of CoManage is expected to remain an issue for the wider astronomical community and a more appropriate solution may be to provide a special purpose service for community management, and e.g. provisioning CoManage from there.

As an alternative, the SURF Research Access Management (SRAM) federated AAI environment was evaluated for the RACC. This is a far more light weight federated AAI solution and found to be very attractive from the user point of view. The web interface natively provides a clean dashboard with an overview of virtual organizations, collaborations (e.g. research projects), and services that the user has access to. As a proof of concept, a few core LOFAR services have been configured to use SRAM as the AAI provider, notably for ASTRON the Tomcat web applications, the LTA portal, and LDAP provisioning and for SURFsara the Spider system. SAML integration was not pursued as the only LOFAR service earlier integrated based on SAML was found to support OIDC as well. Figure 8 shows the user landing page of SRAM.



My memberships

ORGANISATIONS 1
- Astron Test Organization 1

COLLABORATIONS 1
- ASTRON CO 1 [astron_org_1]

SERVICES 3
- EDUteams test service
- Astron test ldap service
- Astron OIDC Service

Fig. 8 – User landing page SRAM

SRAM will require further configuration and battle-testing before being ready for operational deployment in the LOFAR community, but the assessment demonstrates that service integration with another AAI provider is in itself very well feasible.

The federated AAI assessment has shown that to support a research community such as LOFAR, it is critical to consider the user experience and provide appropriate support for setting up and managing collaborations in science projects allowing for a high level of self-supported organisation by service providers as well as researchers. The federated AAI setup investigated for the LOFAR services proves that integration of all key LOFAR services with federated AAI can be realized through a combination of OIDC, SAML (although not required), and public SSH key provisioning. All integration efforts have been documented and an implementation plan is available to follow-up the migration of LOFAR services to federated AAI. The experience will be translated into an update of the RACC requirements for federated AAI based on which ASTRON will select the best matching provider and implement a production AAI system for the LOFAR community to be realized in 2021.

2.11.2.2 Portable data processing workflows for LOFAR data processing

Applicability of the Common Workflow Language to the domain of LOFAR data processing had already been explored prior to EOSC-hub in the EOSCpilot project. The work was continued in EOSC-hub with a rigorous evaluation in comparison with existing processing frameworks of the LOFAR community. The results were presented to the community of software developers. It was concluded that porting of production processing pipelines to CWL would be the preferred way forward as it would provide a standard based and well-supported environment that will speed up pipeline development and will improve pipeline reusability. An example pre-processing pipeline from the LOFAR community is presented in figure 9.

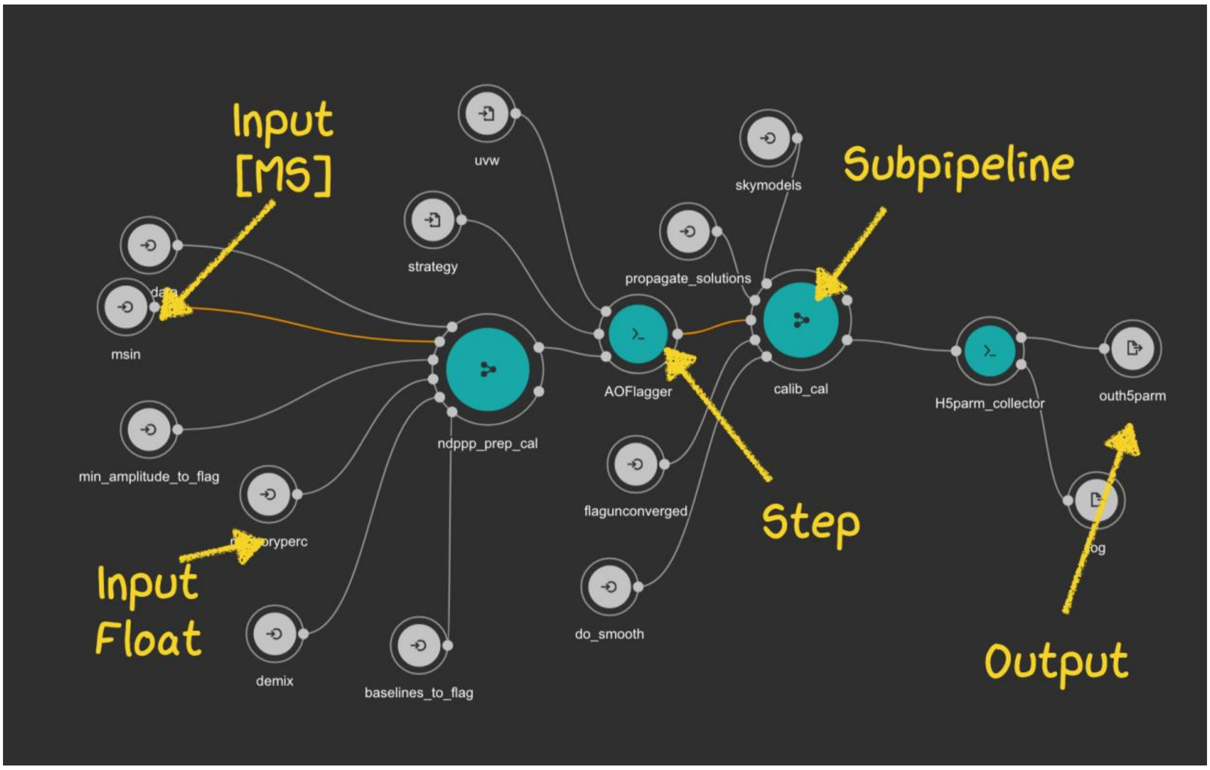


Fig. 9 – Example pre-processing pipeline from the LOFAR community

The community was engaged for selection of a pipeline that would be developed and integrated into a start-to-end LOFAR data processing workflow to be demonstrated for the EOSC-hub project. The selected pipeline is the PreFactor (https://github.com/lofar-astron/prefactor) which is the main pipeline in use for direction independent calibration of LOFAR imaging observation. This pipeline was at the basis of the LOFAR Two-metre Sky Survey First Data Release (doi:10.1051/0004-6361/201833559) and provides an essential step in the generation of science ready data. A library of common CWL steps for LOFAR data processing has been developed and together with the developer of PreFactor it has been ported to CWL (Mancini et al, proc. ADASS 2020).

The Pipeline software was packaged as a Singularity container then was deployed on the SURFsara Spider cluster as the central component for the LOFAR data processing service. The diagram below provides an overview of the architecture of this service. Seven steps have been defined and developed for the RACC:

1. Data is staged within the dCache storage environment which is used at each LOFAR LTA site to ensure files are available for direct retrieval. For this, a new stager service has been developed that provides a REST API callable from the workflow. Since not all LTA sites yet fully support WEBDAV and the dCache REST API, support for the older gfal interface library is implemented as well.
2. Data is retrieved from dCache to the compute cluster (Spider) using the WEBDAV data interface of dCache.
3. The Workflow manager executes the CWL pipeline on Spider, distributing jobs over nodes using the AGLOW (ref.) system or directly using Toil (ref.) The pipeline stores the final data

products in a separate area before it is archived. This allows for evaluation of the results and curation, as necessary. The workflow collects the relevant metadata for later use in the archiving step.

4. A (currently manual) step triggers the generation of PID's and preparation of metadata and datasets for registration in the science data repository (SDR or B2SHARE, see section on FAIR data sharing). The data itself remains on the dCache storage; references to the data endpoints are included in the registration as data links.
5. Data collections can be selected for registration in the Virtual Observatory server hosted at ASTRON. This server exposes data products through standard community protocols developed by the International Virtual Observatory Alliance (IVOA).
6. Once a VO collection is published, it is registered in the central IVOA catalogue and the data becomes discoverable by the community through various VO-enabled tools and web interfaces.
7. Published data collections are automatically harvested by the B2FIND service and become publicly discoverable in the European Open Science Cloud. (This functionality was not developed specifically for RACC so any collection published in the VO will be harvested).

At the time of writing this report, step 4 of the workflow is still under development but aimed to be ready for demonstration first half 2021.

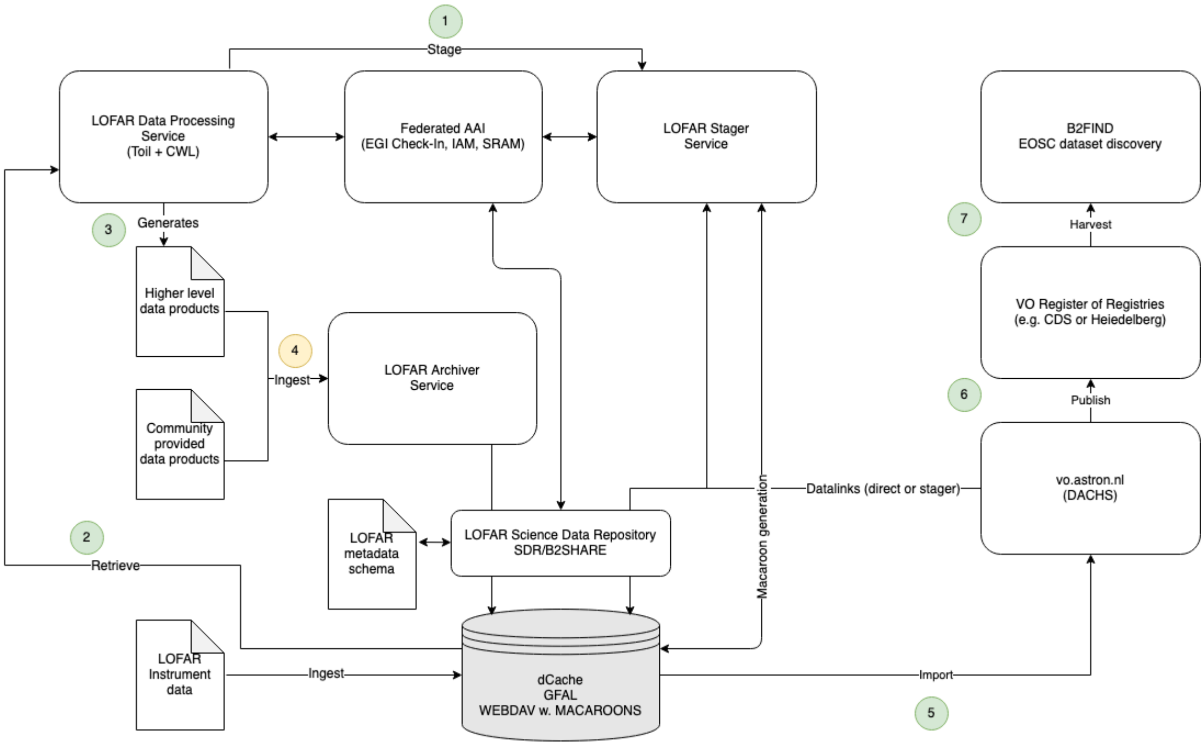

1
Stage
LOFAR Data Processing Service (Toil + CWL)
Federated AAI (EGI Check-In, IAM, SRAM)
LOFAR Stager Service
B2FIND EOSC dataset discovery
7
Harvest
3
Generates
Higher level data products
VO Register of Registries (e.g. CDS or Heiedelberg)
4
Ingest
LOFAR Archiver Service
6
Publish
Community provided data products
Datalinks (direct or stager)
vo.astron.nl (DACHS)
2
Retrieve
LOFAR metadata schema
LOFAR Science Data Repository SDR/B2SHARE
Macaroon generation
LOFAR Instrument data
Ingest
dCache GFAL WEBDAV w. MACAROONS
Import
5

Fig. 10 - LOFAR data processing workflow. Green: realized; Yellow: under development

On the infrastructure side, support was given by SURFsara to:

- Set up LOFAR pipelines and collaborative project spaces within an optimized cloud technology based-batch processing platform (i.e., Spider)

- Set up the LOFAR workflow engine 'AGLOW' on Spider and enhance it to be able to use it with Slurm-based batch processing
- Implement and test new pipeline frameworks based on CWL/Toil on Spider
- Investigate, Implement, test and improve data retrieval and transfer and distribution methods (e.g., Advanced dCache API, object store) for LOFAR data retrieval using bearer tokens, API's and scripts)
- Investigate dedicated private node setups within the SURF cloud for LOFAR customized services such as hosting AGLOW.

The other LTA sites supported the effort through improvement of the local data access capabilities and supporting the local deployment and execution of LOFAR processing workflows.

Various components have been developed to support the described workflow and these have been collected in a public GIT software repository hosted at https://git.astron.nl/eosc. Some of the components have already been applied in related activities. For example to support the metadata collection for publishing Apertif imaging DR1[23].

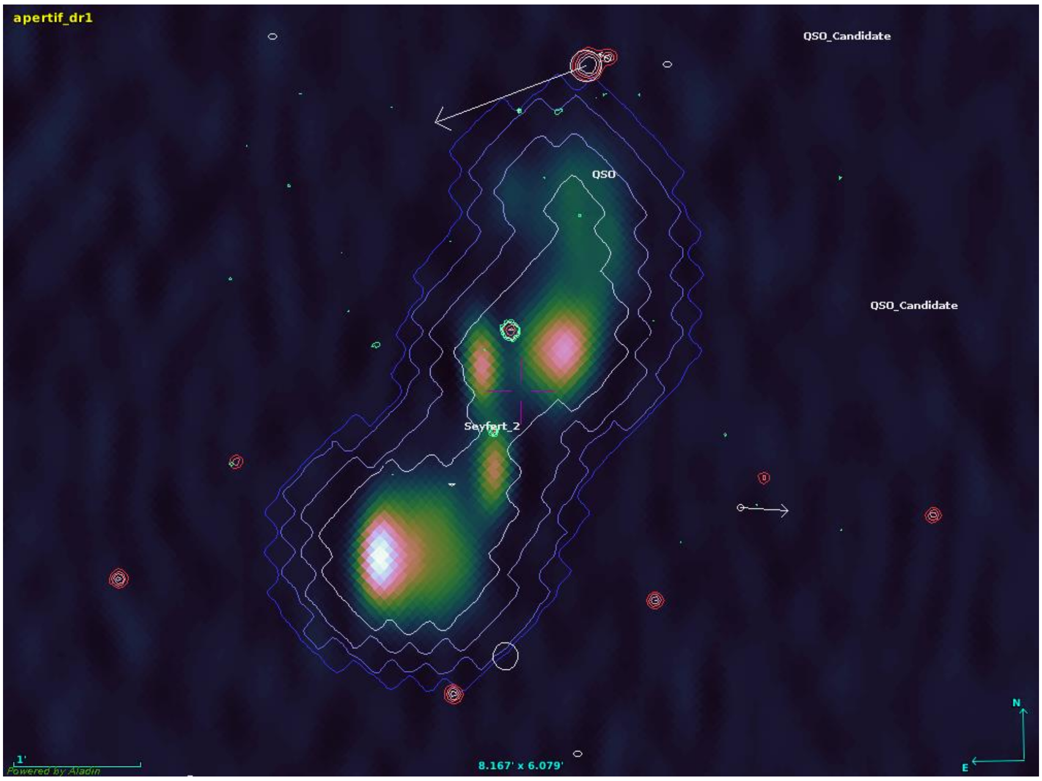


Fig. 11 - Aperitif DR1 image of 2MASX J13284845+2752280 visualized using the Aladin desktop VO tool with NVSS contour overlay (courtesy M. Mancini).

Throughout the process, the workflow has been assessed and tuned in close collaboration with members of the LOFAR Survey community, ensuring the quality of the results produced by it.

[23] https://www.astron.nl/telescopes/wsrt-apertif/apertif-dr1-documentation/

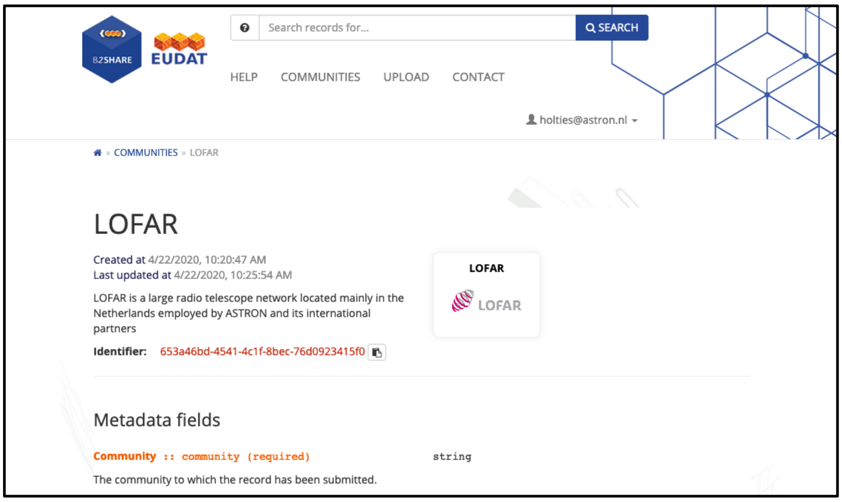
Moreover, the LOFAR processing case has been provided as a use case for the EOSC-hub Business Models Study (EOSC-hub Briefing Paper –Provision of Cross-Border Services, Oct 2020[24]).

2.11.2.3 Improving FAIR data access to high level science data products

In the area of FAIR data sharing, the RACC has created an initial metadata schema for describing LOFAR data products. The schema is compatible with the IVOA obscore model[25], extended with specific attributes which allow the description of interferometric data such as generated by the LOFAR instrument. The schema has been implemented in B2SHARE and in the SURF Data Repository (SDR). Test instances of these repositories have been used to demonstrate the registration of LOFAR data collections such as generated by the LOFAR processing workflow or e.g. delivered by science teams in the community. SURFsara has supported the effort by enhancing the capabilities of the SURF data repository, notably by working on the automation of data object ingests.

Also ASTRON has acquired its own ePIC handle prefix (21.12136) for assigning PIDs to archived data products, data collections and other publications from amongst others LOFAR and APERTIF.

Figures 12 & 13 show LOFAR data product entries in the B2SHARE catalogue, and the LOFAR repository entry point within the SURFsara Data Repository.



LOFAR

Created at 4/22/2020, 10:20:47 AM
Last updated at 4/22/2020, 10:25:54 AM

LOFAR is a large radio telescope network located mainly in the Netherlands employed by ASTRON and its international partners

Identifier: 653a46bd-4541-4c1f-8bec-76d0923415f0

Metadata fields

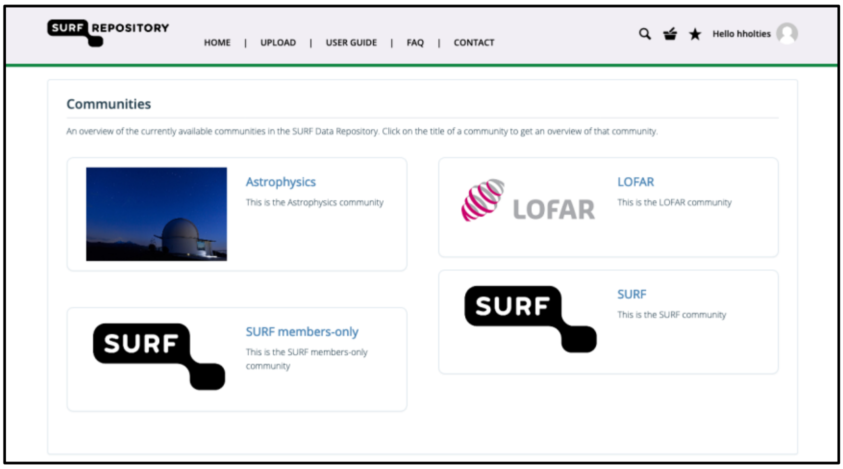
Community :: community (required) string

The community to which the record has been submitted.

Fig. 12 - LOFAR data product entries in the B2SHARE catalogue

[24] https://www.eosc-hub.eu/sites/default/files/EOSC-hub%20Briefing%20Paper%20-%20Provision%20of%20Cross-Border%20Services%20-%20For%20Consultation.pdf

[25] https://www.ivoa.net/documents/ObsCore/20170509/REC-ObsCore-v1.1-20170509.pdf



Communities

An overview of the currently available communities in the SURF Data Repository. Click on the title of a community to get an overview of that community.

Astrophysics

This is the Astrophysics community

LOFAR

This is the LOFAR community

SURF

This is the SURF community

SURF members-only

This is the SURF members-only community

Fig. 13 - LOFAR repository entry point within the SURFsara Data Repository

A metadata extraction software component has been developed for collecting metadata from radio astronomical data products in accordance with the LOFAR schema. This component is embedded in the LOFAR processing workflow as a CWL step.

An archiving component is currently under development as the final step for the beginning to end LOFAR processing workflow. This component will allow a user to complete the metadata with e.g. information of the team responsible for the generation of the data products and to annotate it with information about the scientific purpose and quality. The registration of a dataset in one or more of the science data repositories (B2SHARE, SDR, IVOA) can thereafter be initiated which, upon successful validation of the provided information, will also generate Persistent Identifiers with an ASTRON prefix for all registered data products as well as the collection.

2.11.3 Impact

The RACC focused on demonstration of key components and did not deliver new services to an operational level such that they could be registered in the EOSC portal. The developed software components are available from https://git.astron.nl/eosc and the activities have been documented on the ASTRON Confluence system.

The services under development have been used internally and have been demonstrated to members of the LOFAR community.

A training session on the LOFAR processing workflow will be included in the upcoming LOFAR data school which is planned for March 2021 and where some 40 users are expected to participate.

2.11.4 Future plans beyond EOSC-hub

The EOSC-hub project has allowed the Radio Astronomy Competence Center to make significant progress towards achieving its ambitions. As a competence center, the primary objective has been to demonstrate how the specific challenges for the data intensive radio astronomy community can be addressed by building on infrastructure services available in the European Open Science Cloud and to explore the applicability of general purpose EOSC data services. The achieved results will be built upon in follow-up projects through which a LOFAR data processing service will be offered to the science community in 2021. In 2020, ASTRON has formed an operational Science Data Center

(SDC) group that will take responsibility for offering data services to the astronomical community. The accompanying SDC Program will oversee the further development of services and deliver them to SDC Operations, thus providing the organizational structure for continued sustainable development and operation.

The first project that will benefit from the achievements for the RACC, and where they will be applied at scale, will be the LOFAR Data Valorization project. This project is funded by NWO-I ASTRON and will retrieve up to 25 PB of low-level data products from the LOFAR LTA in 2021 to perform data curation and compression and generate some 9 PB of compressed higher level data products. The latter are a good starting point for further analysis and are aimed to attract a broader community of researchers to benefit from the valuable data contained in the LOFAR LTA.

The LDV project will bring the LOFAR processing workflow to a sufficiently high technical readiness level to start offering it as a user service by the end of 2021. The user service will allow astronomers to request processing of fully direction independent calibrated data, where the calibration is performed on the Spider compute cluster hosted at SURFsara and the calibration is performed using an updated version of the PreFactor CWL pipeline. This processing service will be offered as a Virtual Access service under the EGI-ACE H2020 project (to start in January 2021). The resulting data products will be registered into a science data repository which will be provided through the DICE project. ASTRON aims to have a federated AAI environment ready in time for operational use by these services and start upgrading existing LOFAR services to use federated AAI.

The new stager service will replace the existing one after having been fully commissioned where it will become part of the LOFAR LTA Virtual Access service which is co-funded under the awarded ORP project. The metadata extraction and archiving components will be used in future data releases for LOFAR and Apertif of which several are planned for 2021.

All results and experience will be fed back to the ESCAPE project and to partners in the ORP project to benefit the development of data services for astronomical instruments other than LOFAR, including the SKA.

2.13 ICOS-eLTER

2.13.1 Initial ambition (in 2018)

The ICOS - eLTER Competence Centre groups two scientific communities:

1. The Integrated Carbon Observation Systems (ICOS) research infrastructure: https://www.icos-ri.eu
2. Long-Term Ecosystem Research in Europe (eLTER) community: http://www.lter-europe.net

The ICOS group integrates and tests generic services from the EOSC-hub portfolio with the ICOS Carbon portal to provide a scalable environment for researchers wishing to monitor and analyse carbon processes. The CC expected to have the first runs of the portal with actual near real-time data performed before the end of 2018 and to start full production or near real-time data in January 2019. (However this is delayed to the future due to both technical and organisational reasons which are beyond the CC. - See details later.) The resulting data is to be automatically ingested and be published in the ICOS Carbon Portal and at the Thematic Center.

The eLTER group aims to develop and share data quality control workflows with the Long-Term Ecosystem Research communities through user-friendly interfaces and on top of scalable compute systems. The group will test suitable common services from EOSC-hub to assess the best fit for the eLTER purposes, then based on the experience arrange the long-term setup at the core eLTER RI sites.

2.14 ICOS

2.14.1 Progress and key results

The purpose of the project is to develop the near-real time data processing of eddy covariance flux data for greenhouse gases into a web enabled cloud module where data flows directly from the field instruments to the ICOS Ecosystem Thematic Centre (ETC) and is processed and reduced from 20 Hz into half hourly flux data inclusive data quality and uncertainty parameters. The system is meant to be very flexible and is to be used by ICOS for its ecosystem eddy covariance data but can also be provided as a service to other infrastructures of scientific users to process their data using the ICOS high quality data processing.

Part of the system is the automatic ingestion of data from a variety of data acquisition systems used by the 84 ICOS ecosystem stations. All instruments produce for each station a 20 Hz signal of the three components of wind speed, temperature and CO2 concentration. Through statistical analysis this is convolved into vertical exchange fluxes of latent and sensible heat flux and the net CO2 exchange flux over a certain averaging period, usually 30 minutes. Every half hour the field instruments or their data loggers send the data accumulated over that period to the ICOS Carbon Portal to the so called ETC Facade where the data is checked against a data checksum and checked on data validity. On the next day, the full set of half-hourly files are concatenated to daily files per station, ready for the statistical data processing. At that moment, each file is ingested in the ICOS data store and minted a Handle PID together with the relevant provenance metadata according to the data object type. Part of the ingestion is the storage of the data object at the EOSC B2SAFE service.

For the data analysis a program was built using R. The software is described in a paper[26,27]. A wrapper software was developed using Ansible and R that harvests the available data file PIDs and their metadata from the ICOS repository using a SparQL query and then runs the processing code to process the files with the found PID identification in a Docker virtual machine. The script takes care of launching docker instances for every station that has new data available. For each station run the script gathers the metadata from the ICOS Ecosystem Thematic Center to set the parameters for the data processing, that describe the station and important features like gap-filling, coordinates etc. The resulting data files that contain the half hourly processed fluxes are then ingested at the data repository as a so called Near Real Time data objects (Level 1) or final quality-controlled data (Level 2) together with the relevant metadata on provenance (Software version, station, instrument, files used etc).

The original ambition was to make use of the EOSC cloud service at CSC, to enable use of the cloud processing also by external users outside ICOS and to provide a user-friendly interface for submission of data for processing and visualisation and download of the end result.

At this moment, the system is partially ready and has been used for processing the ICOS Level 2 release of 2020 (ICOS Research Infrastructure 2020. Ecosystem eddy covariance final quality (L2) flux product in ETC-Archive format - release 2020-1 (Version 1.0). ICOS ERIC - Carbon Portal. https://doi.org/10.18160/ABWE-HMV4), with the notion that the docker machines have been deployed at the ICOS data portal server fsicos3. The connection with B2SAFE has been established, so all raw data and processed data is stored and duplicated at this trusted repository. This dataset of the 16 labelled ICOS ecosystem stations has now been downloaded 60 times and most files have been previewed about 90 times.

In principle launching of the calculations in the EGI and CSC cloud instead of on the ICOS server would be trivial, except for the high demand on storage and memory. By launching the calculations on our local server that has local copies of the relevant input files we can ensure the required performance and avoid long delays in the transfer of data files. Our current server has 128 cores and 100 TB storage which still scales well with the current needs.

Delays in the implementation were partly caused by the COVID-19 situation which complicated the work at the ICOS Ecosystem Thematic Center (ETC) in Italy considerably, for example for the training of the personnel in Italy to work with SparQL queries, Unix (ETC is Windows Server and SQL centric), the docker components and the parallel processing of R code. About ⅓ of the available budget, reserved for the implementation of the web-based interface, has not been used.

[26] Sabatini et al., 2018: Eddy covariance raw data processing for CO2 and energy fluxes calculation at ICOS ecosystem stations, https://www.research-collection.ethz.ch/handle/20.500.11850/313355

[27] Vitale et al., 2020: A robust data cleaning procedure for eddy covariance flux measurements, https://bg.copernicus.org/articles/17/1367/2020/

2.14.2 Lessons learnt

Despite the promises, data processing in the cloud when it concerns a large number of relatively large data files and memory intensive processing is a challenge when both data transfer and CPU memory limit performance and/or increase costs. And as always developing robust data processing of a complicated and diverse set of stations with high demand on scientific quality is a difficult task that exceeds any planning easily.

2.14.3 Impact

The developed service has been used to generate the Level 2 ICOS final quality-controlled dataset release for 2020.

As the web-based service for external users is not ready yet, external users could not be trained. ICOS CP has trained the staff at ICOS ETC in use of the command line-based services.

2.14.4 Future plans beyond EOSC-hub

The system developed will be used by ICOS, the global fluxnet database and be open for the user community. There is already strong interest from the South-African ecosystem infrastructure SAEON into the data processing system.

2.15 eLTER

2.15.1 Progress and key results

Throughout the EOSC-hub project, eLTER developed a prototypical Web service providing a workflow for outlier analysis in time-series data fetched as datasets from cloud-based repositories or dynamically via OGC Sensor Observation Service (OGC SOS). Due to the availability of existing implementations of different outlier detection methods of varying complexity, the workflow was developed using the R statistical programming language. The flexibility of this development environment also allowed to implement the data retrieval functionality there, using standard Web requests for data retrieval from file-based repositories and a dedicated library (SOS4R) for retrieval from OGC SOS. The complete R workflow, returning the outlier-annotated original time-series data as well as more in-depth information about the individual detection runs, was encapsulated behind a Rest API implemented using a Python Flask Web server automatically created as a stub from a formal OpenAPI specification. The Web service was developed and deployed in a Virtual Machine on the EGI Cloud (using the access.egi.eu resource pool that is available for piloting). The Web service is integrated with B2DROP as a file-based data repository. EGI Notebooks were used as the access layer to demonstrate the functionality of the service.

The setup successfully demonstrated the feasibility of providing cloud-based Web services for outlier analysis in time-series data, fetching source data directly from other cloud-based sources without introducing the need to intermediately store them on user machines. This paved the way to providing generic data quality assurance services suitable for a wide variety of time-series data across many different domains, operated and orchestrated using EOSC based services.

The service prototype was internally validated using different test runs operating on data fetched from different file-based repositories as well as SOS services. Calibration of the methods specification and validation of the performance of the statistical methods was done visually by producing figures of the studied time-series with the identified outliers highlighted in color. Validation with time-series complemented with artificial outliers was prepared and test runs were conducted.

2.15.2 Lessons learnt

The evaluation of the initial prototype raised a number of additional requirements. On the one hand, the sequential applications of multiple outlier detection methods consumed significant resources especially for longer time-series, calling for efficient parallelization of related invocations, which should be easily achievable using existing EGI cloud resources. On the other hand, especially the data retrieval via SOS revealed quite a variety of existing server configurations, resulting in the need to apply source-specific parameter combinations for successful data retrieval and calling for community based processes to establish dedicated configuration profiles. Related to the statistical methods used, the evaluation showed that the performance differed substantially between the applied environmental parameters due to different characteristics of these time-series. Thus, testing and tuning of the input parameters of the methods and the service is essential before use.

2.15.3 Impact

Being of prototypical nature, the outlier detection service has not been registered in the EOSC portal. The service was only used in internal test runs but will further be exploited in the eLTER PLUS project. The prototype was presented to a large audience at two different conferences (EGU2020 and Geospatial Sensing 2020) which were held online due to the COVID-19 crisis. The presentations raised significant interest in the service and opened new perspectives for further development.

2.15.4 Future plans beyond EOSC-hub

The development of the prototype into a fully operational and functional service, which should include technical improvements with respect to parallelization as well as community based processes such as finding a common SOS profile, will be continued throughout the eLTER PLUS project. Furthermore, the performance of the service (particularly the environmental parameters and statistical methods used) should be tested by running the service with time-series complemented with artificial outliers of different magnitude and frequency.

2.17 Disaster Mitigation Competence Centre plus (DMCC+)

2.17.1 Initial Ambition (2018)

Disaster mitigation is one of the key sustainability issues of most Asian countries as Asia region suffers the most total damage (covers losses of human lives and economy etc.) from disasters in comparison to other regions in the world in the past three decades. DMCC+ aims to support capacity development on hazard risk analysis based on the 'deeper understanding' approach over the regional open science cloud platform with the collaboration of 10 Asia countries and 2 European partners. The deeper understanding approach aims to discover the root causes and physical processes of target events. And accurate simulations of the whole lifecycle of target events are developed accordingly to quantify the risks. Through such regional collaborations, the distributed cloud regional infrastructure is enhanced, based on EGI/EOSC-hub compatible technologies. Case study, simulation portal and knowledge base constitute the framework for disaster mitigation capacity development Case study drives advancement of science, technology, simulation portal/applications and collaboration. Partners are able to reproduce the hazard events numerically through simulation portals and carry out risk analysis. All the observation data, input parameters, simulation tools, event facts and publications etc will be compiled into the knowledge base. Based on lessons learned from case studies, partners are able to make use of numerical simulations for better risk analysis on other events and conducting their own case studies. Simulation facilities and knowledge base will be growing and strengthened progressively with more case studies.

2.17.2 Progress and key results

2.17.2.1 Case studies

Case studies of this collaboration aimed to quantify risks and reduce at least one of the dominating factors of vulnerability, exposure, and hazards based on the up-to-date knowledge of physics behind the disaster. Activities in this are included:

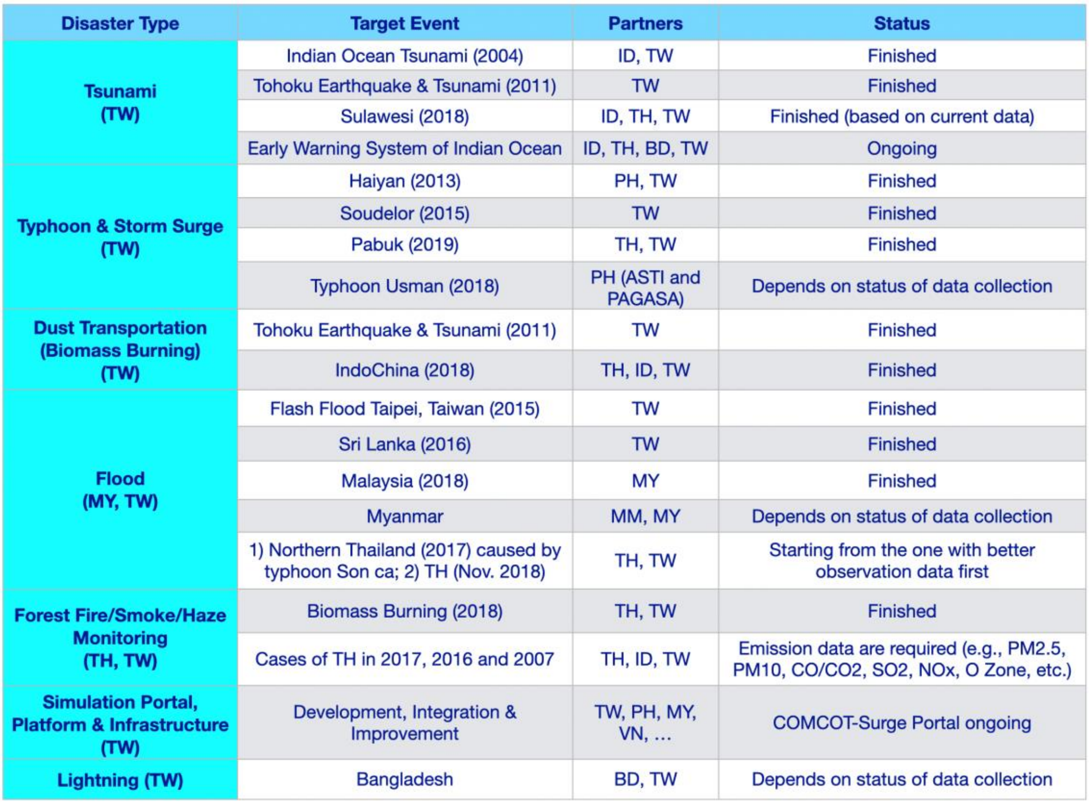
1. Translating evolving scientific advancement into accurate numerical simulations;
2. Understanding the trends of changes of hazard impacts;
3. Developing risk analysis and mitigation capability;
4. Building up flexible and dynamic collaboration models of all parties;
5. Improving the distributed cloud infrastructure to support digital simulations.

18 case studies of 6 types of hazards in almost every partner country has been conducted. 13 of them had finished. The hazard types cover tsunami, typhoon and storm surge, dust transportation, flood, forest fire/smoke/haze monitoring, and lightning. For each type of hazard, the leading scientists or scientist group has been assigned or formed, to conduct the deeper understanding of events and develop and validate the simulation processes. Now partners such as Philippines, Thailand, Vietnam can conduct their own case studies based on the deeper understanding approach and contribute to the regional collaboration framework. From the routine meetings, we support and encourage partners to bring their case studies and users to work with us, in addition to sharing their requirements, case study experiences etc.

Three case studies results are demonstrated here as examples:

1. 2018 Sulawesi Tsunami: The case study was conducted because of the complexity in source identification and characterization. According to the scenario studies, we concluded tentatively that there are two tsunami sources in this event. A strike-slip earthquake tsunami dominates the impact outside the Palu Bay. Inside the Palu Bay, landslides are one possible source. All these analyses are able to be simulated by iCOMCOT portal.
2. The dispersion of aerosols from biomass burning in Indochina in springtime of 2018: The simulation captures the route and sampled concentration of the dust thousands km long-distance transportation. This case also had been verified by data collected from ground-based gauges by Thailand.
3. The typhoon that ran through Thailand from SCS to Andaman sea in early Jan 2019. Our simulation explained how the wind speed, air pressure, precipitation and temperature contributed to the typhoon generation process. We used the WRF application for the simulation.

Case study of super typhoon Haiyan is a representative example of a deeper understanding approach. Typhoon Haiyan caused >6,300 life losses in the Philippines in 2013. Most of the life losses were resulted by storm surge induced by the typhoon. In order to capture the impact of storm surge timely and accurately, we combined the atmospheric model and oceanic model together to simulate the storm surge effects directly from the evolution of the typhoon. The other challenge is that it is very difficult to precisely simulate the lowest pressure and highest wind speed of a super typhoon. In this case study, the simulation model was revised based on the novel studies of eyewall physics. A status overview of the use cases is summarised in the following table.



Disaster Type	Target Event	Partners	Status
Tsunami (TW)	Indian Ocean Tsunami (2004)	ID, TW	Finished
	Tohoku Earthquake & Tsunami (2011)	TW	Finished
	Sulawesi (2018)	ID, TH, TW	Finished (based on current data)
	Early Warning System of Indian Ocean	ID, TH, BD, TW	Ongoing
Typhoon & Storm Surge (TW)	Haiyan (2013)	PH, TW	Finished
	Soudelor (2015)	TW	Finished
	Pabuk (2019)	TH, TW	Finished
	Typhoon Usman (2018)	PH (ASTI and PAGASA)	Depends on status of data collection
Dust Transportation (Biomass Burning) (TW)	Tohoku Earthquake & Tsunami (2011)	TW	Finished
	IndoChina (2018)	TH, ID, TW	Finished
Flood (MY, TW)	Flash Flood Taipei, Taiwan (2015)	TW	Finished
	Sri Lanka (2016)	TW	Finished
	Malaysia (2018)	MY	Finished
	Myanmar	MM, MY	Depends on status of data collection
	1) Northern Thailand (2017) caused by typhoon Son ca; 2) TH (Nov. 2018)	TH, TW	Starting from the one with better observation data first
Forest Fire/Smoke/Haze Monitoring (TH, TW)	Biomass Burning (2018)	TH, TW	Finished
	Cases of TH in 2017, 2016 and 2007	TH, ID, TW	Emission data are required (e.g., PM2.5, PM10, CO/CO2, SO2, NOx, O Zone, etc.)
Simulation Portal, Platform & Infrastructure (TW)	Development, Integration & Improvement	TW, PH, MY, VN, …	COMCOT-Surge Portal ongoing
Lightning (TW)	Bangladesh	BD, TW	Depends on status of data collection

2.17.2.2 Simulation portals

The CC continued to use the iCOMCOT portal that was developed in the EGI-Engage project before EOSC-hub. During EOSC-hub the portal[28] (iCOMCOT) started onboarding to the EOSC Marketplace. The process was suspected because iCOMCOT is in a long shutdown to be migrated to a container-based cloud setup to improve availability-reliability. The plan is to reopen iCOMCOT services in January 2021 and resume the onboarding in EOSC then.

The gWRF weather simulation portal, which was used in the first part of the CC, was removed from production to upgrade the software and to re-package it into a container (MPI application).

2.17.2.3 Training & dissemination

The CC leveraged existing collaborations and large-scale events, such as APAN, Asi@Connect project meetings and the International Symposium on Grids and Cloud (ISGC) to maintain collaboration among CC members, to engage with new partners, with local user communities, to host training events and masterclasses and to disseminate our work. 4 training events (in New Zealand, Mianmar, Bangladesh, Malaysia), 2 workshops (at the ISGCs in Taipei) and 4 collaboration meetings (Singapore, Malaysia, Korea, and a virtual one) had been hosted to enhance the regional collaborations on case studies and experiences sharing.

2.17.3 Impact

Our iCOMCOT simulation portal served over 40 researchers from 8 Asian countries. The training events were attended by over 200 participants.

Myanmar and Bangladesh joined the collaboration as new partners. Nepal has brought case studies to work with DMCC+ in 2019. Local communities in the Philippines, Malaysia, Bangladesh have also joined the case studies of DMCC+.

Publications: 4 academic papers had been published based on DMCC+ case studies:

- Yen, E., Lin, S., Wu, T.R., Tsai, Y.L., and Chung, M.J., (2020), Knowledge- Building Approach for Tsunami Impact Analysis Aided by Citizen Science, Frontiers in Earth Science-Solid Earth Geophysics.
- Kueh, M.-T., Chen, W.-M., Sheng, Y.-F., Lin, S. C., Wu, T.-R., Yen, E., Tsai, Y.-L., and Lin, C.-Y. (2019): Impacts of Horizontal Resolution and Air-Sea Flux Parameterization on the Intensity and Structure of simulated Typhoon Haiyan (2013), Nat. Hazards Earth Syst. Sci. Discuss., https://doi.org/10.5194/nhess- 2018-333, 2019
- Yen E., Chiang J. (2018) Development of Open Collaboration Framework for Disaster Mitigation. In: Bungartz HJ., Kranzlmüller D., Weinberg V., Weismüller J., Wohlgemuth V. (eds) Advances and New Trends in Environmental Informatics. Progress in IS. Springer, Cham
- Yen E. ; Lin SC ; Tsai YL; Wu TR; Lin CY, (2018) Open Application Framework for Disaster Mitigation Based on Deeper Understanding Approach, 2018 5th International Conference on Information and Communication Technologies for Disaster Management (ICT-DM)

[28] https://icomcot.twgrid.org

2.17.4 Future plans beyond EOSC-hub

The regional collaborations on disaster mitigation in Asia extend beyond EOSC-hub, based on the foundations laid by the DMCC+, and based on the ‘deeper understanding’ approach. More case studies about various types of hazards and from different countries will be commenced. They will focus on local interests and will be led by local communities within the partner countries. More Asian partners will join these collaborations, bringing their case studies and sharing their requirements or experiences.

Additionally, DMCC+ has been piloting applications on impact analysis on agriculture, in collaboration with regional agricultural efforts. Extension to space-based resource federation and applications will be prototyped in the coming years in the context of the EGI-ACE project. Capacity building for dealing with more complex scenarios such as multi-hazards and compound hazards will be included in future case studies.

All these activities will be integrated by an open collaboration framework over the distributed cloud infrastructure, which is compatible with the future EOSC Compute Platform. The primary objective of this framework is to capture knowledge learned from all the case studies and support knowledge sharing, reuse and repurposing for broader application possibilities based on the European Open Science Cloud paradigm.

3 Conclusions and observations

This section articulates some of the common or major ‘lessons learnt’ from the overall CC activities.

1. Despite there were delays with the start in a few of the communities (EISCAT_3D, Radioastronomy, ICOS-eLTER, SeaDataNet, eLTER), this did not prevent these CCs reach their original objectives.
2. Federated identity management was the most important common topic across the CCs. Nearly every CC had a related activity, and they all successfully enabled federated login for their users via the EOSC-compliant AAI proxies (EGI Check-in, B2ACCESS, IAM, eduTeams).
3. Pre-caching of datasets at compute clouds, and downloading of chunks of data into clouds only when and where they are needed was identified as the preferred way of data distribution in several CCs, vs. the ‘replicate everything to every partner cloud’ model. (ELIXIR, SeaDataNet, Fusion) EGI DataHub was the service for implementation where implementation took place for such cases (SeaDataNet, Fusion).
4. Cloud providers work with different billing units and cost models. It would be a very high effort to bring all such providers to a common billing unit to enable cross-site billing (with either virtual tokens or real funding). This effort is very difficult to justify if the reimbursement model is not known upfront. (ELIXIR)
5. Critical to consider the user experience to reach a system that is attractive for a high number of users/access projects. E.g. use the same authentication method, use similar and ‘scientist friendly’ look-and-feel on the GUI of the various tools that underpin an overall, typical research workflow. (RACC, Fusion)
6. Despite the promises, data processing in the cloud when it concerns a large number of relatively large data files and memory intensive processing is a challenge when both data transfer and CPU memory limit performance and/or increase costs. (ICOS)
7. Demonstrated the feasibility of providing cloud-based Web services for data cleaning, alignment, paving the way to online analytics for time-series data across many different domains. (eLTER)
8. Adaptable code, modular programming, containerization are key elements to achieve portable, flexible and maintainable applications and infrastructures. This is exceptionally important to stay resilient in case of staff turnover, and with development programmes spanning multiple projects. (ELIXIR, Euro-Argo, Radioastronomy)
9. Jupyter notebook (JupyterHub) as a user access and data analysis environment is prominent, and successful when a centrally operated JupyterHub can serve the whole international community. However achieving a homogeneous setup with multiple JupyterHubs deployed in different countries is challenging, due to local constraint and established practices at compute centres. (Euro-Argo, EISCAT_3D, EPOS-ORFEUS)
10. The overall support for RIs was found successful and sufficient in the project. The CCs integrated RI representatives, e-infrastructure and software experts into small projects that received additional support from generic service teams from WP10. Face-to-face and online meetings were found effective for cross-CC and cross-WP interaction, online tools (Confluence and Jira) brought little added value in this area.

Appendix I. Service assessment & adoption matrix

Summary of the service assessment and integration work of the CCs.

		Community	ELIXIR	Fusion	Euro-Argo	SeaData Net	EISCAT_3D	EPOS-ORFEUS	Radio astronomy	ICOS	e-LTER	DMCC+
		Services brought into the EOSC Portal	Institutional clouds (CESNET, CSC, EBI)	PROMINENCE	Argo Floats Data Discovery		EISCAT Data Access Portal		LOFAR science products (to be ready in 2021)	ICOS Portal (in onboarding pipeline)		iCOMCOT (to be ready in 2021)
Services used from EOSC-hub	**Compute**	DODAS		▨								
		EGI Cloud		🟩			▨			▨	▨	
		EGI Workload Manager					🟩					
		EGI HTC					▨		🟩			🟩
		Jupyter (EGI Notebooks and Custom)			🟩		▨	🟩				
	Data	EGI DataHub		🟩		▨						
		B2SAFE		▨				▨		🟩		
		B2Stage						▨				
		B2Handle						▨				
		B2Share							🟩			
		B2Drop			▨			▨			▨	
		B2Find					▨		🟩			
	AAI	Check-in	🟩	▨			🟩	▨	▨			
		B2ACCESS	🟩	▨				🟩				
		IAM		🟩								
Services used from outside the project	dCache								🟩			
	Institutional clouds		🟩		🟩		🟩	▨		🟩		▨
	eduTeams			▨								
	Cassandra, Elasticsearch				🟩							
	SURF tools (Research Access Management; Data Repository)								🟩			
	Containers			🟩			🟩		🟩			

Colour codes:

GREY: technology was validated by the community but there is no immediate use in thematic service after EOSC-hub.
GREEN: technology was positively evaluated by the community and integrated into thematic EOSC service in EOSC.