D8.1 Report on progress, achievements and plans of the Competence Centres

Lead Partner:	EGI Foundation
Version:	1
Status:	Under EC review
Dissemination Level:	Public
Document Link:	https://documents.egi.eu/document/3485

Deliverable Abstract
WP8 includes eight Competence Centres (CCs) that work on establishing infrastructures to support users cope with the data deluge, with the challenges of various compute intensive data analysis scenarios. This document provides a summary of the use cases that drive the CC activities and informs readers about progress with their implementation at half-time of the project. The 8 CCs experiment with 15 common services from the EOSC-hub catalogue, and with 5 services/technologies from outside the project. 3 CCs reached mature integration between common services and their community-specific services while the other 5 are still in intensive integration and technology assessment.

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	8/07/2019
Moderated by:	Małgorzata Krakowian	EGI Foundation / WP1	
Reviewed by:	Niklas Blomberg Ognjen Prnjat	EMBL-EBI / EOSC-Life GRNET / WP4-5-6-7	21/Jun/2019 03/Jul/2019
Approved by:	AMB		23/07/2019

DOCUMENT LOG

Issue	***Date***	***Comment***	***Author***
v.0.1	19/Jun/2019	Full draft sent for external review	Gergely Sipos Steven Newhouse Susheel Varma Shaun de Witt Thierry Carval Ingemar Häggström Carl-Fredrik Enell Luca Trani Javier Quinteros Hanno Holties Rob van der Meer Alex Vermeulen Margareta Hellström Eric Yen and Simon Lin
FINAL	08/Jul/2019	Updated draft after external review	Gergely Sipos

TERMINOLOGY

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

Contents

Executive summary

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

The work started within each CC drafting plans concerning their envisaged use of the EOSC-hub common services based on experiences from predecessor projects (EGI-Engage, EUDAT2020, EOSCpilot). These plans were shared across the CCs during the kick-off meeting, during the EOSC-hub week, the monthly CC coordinators teleconferences, and by the beginning of PY2 evolved into documents within the Wiki and into technical requirements in JIRA communicated to service developers (The Wiki pages and the technical requirements together formed M8.1 in March 2019).

The plans and progress shows that the 8 CCs experiment with 15 common services from the EOSC-hub catalogue, and with 5 services/technologies from outside the project (See Appendix 1 for a table overview of which CC adopts which service). 3 CCs reached mature integration between common services and their community-specific services and are ready for opening up their setups for external users: Marine (ARGO data discovery platform), EISCAT_3D (EISCAT Portal) and ICOS-eLTER (ICOS Carbon Portal). ARGO and EISCAT_3D will launch their service as Thematic Services via the EOSC Portal during the summer and will support operation with Virtual Access in WP13. While the ARGO data discovery platform will be fully open access, the EISCAT Portal will work with access approvals using the EOSC-hub Marketplace and AAI systems. Although the third service, the ICOS Carbon Portal also reached mature setup, due to delays in the ICOS RI data production activities (outside EOSC-hub), the service cannot be opened yet. Because of this, and because of the hiring delay in the eLTER part of this CC the ICOS-eLTER CC is going to be extended in its lifetime until the end of the project.

The other CCs (ELIXIR, Fusion, SeaDataNet in Marine, EPOS-ORFEUS, Radioastronomy, Disaster Mitigation) are progressing with their technology/service evaluation and application integration according to plans, with adopting Cloud, AAI, DataHub, B2SAFE, Notebooks, B2SHARE, B2Find services. More concrete results are expected in Q3-4 of 2019.

The document closes with a section of 5 observations concerning the work done so far by the CCs, and the work still to be done in the remaining 18 months.

1 Progress, achievements and plans by CCs

1.1 ELIXIR

1.1.1 Ambition

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.

1.1.2 User stories

No.	User stories - ELIXIR
US1	ELIXIR wants to establish a federation of cloud sites, each providing storage and compute capacity for researchers. The federated clouds should be connected to a data replication service (Reference Data Set Distribution Service with the ELIXIR terminology - RDSDS) that enables ELIXIR to stage 'ELIXIR Core Data Resources' to the cloud sites on-demand. As a result, the cloud sites become data hosting nodes which are equipped with CPUs/GPUs and are suited for large-scale data analysis and analytics. Centrally provided and curated datasets can ensure high-quality research in any of the partner states/regions. Researchers can go to their 'local' ELIXIR cloud provider, choose an already pre-staged ELIXIR dataset or request the staging of an ELIXIR dataset, choose an application of their choice (from a VM catalogue or container catalogue), maybe upload some additional data and then perform data analysis/analytics. CC members envisage that different conditions of access will apply at the different cloud sites. However, it is expected that the cloud compute resources would be free at the point of use for local researchers and some form of pay-for-use conditions would apply for others. This is because national funding agencies who fund the compute centres require them to serve national research. (The CC will explore possible business models in PY3) The replication of community assets to national cloud providers maximises the utilisation of national funding and lowers the total cost of access for researchers.

	The services in the setup should recognise users via their ELIXIR identity, therefore ELIXIR AAI (Life science AAI) should be integrated with the RDSDS as well as with the national clouds.
US2	The cloud federation can be also equipped with a 'container replication and orchestration service' that enables application providers to deploy containerised community/reference applications to any of the federated cloud sites, and users to instantiate and use the applications on those sites. Having a centrally managed or self-managed container orchestration service will allow users who do not have access to cloud resources of their own to deploy containerised workflows co-located with existing datasets in cloud locations. The services in the setup should recognise users via their ELIXIR identity, therefore ELIXIR AAI (Life science AAI) should be integrated with the RDSDS as well as with the national clouds.

1.1.3 EOSC-hub services assessed, integrated

EGI Cloud; EGI Data Transfer; EGI Check-in

1.1.4 Progress, status, plans

The activities carried out by the ELIXIR-CC in the first eighteen months of the project focused mainly on developing and deploying a Reference Data Set Distribution Service (RDSDS) and develop alignment between ELIXIR and EOSC AAI. Despite personnel delays, the CC managed to complete the development and deployment of the RDSDS service and a number of ELIXIR reference datasets (1000Genomes & MMG) have been made available via the service. Additional datasets (Metabolights, Pride, ArrayExpress and ENA) are being processing for inclusion within the RDSDS index. The period also saw the improved alignment between ELIXIR and EOSC AAI via technical developments within EOSC-hub and strategic developments within the AARC2 project. A particular technical highlight over the last period has been the ELIXIR AAI integration with two ELIXIR OpenStack providers (De.NBI & EMBL-EBI) via OIDC which in turn has technically enabled EOSC-compatible EGI Check-in users to access EOSC-compatible compute services via ELIXIR AAI. The CC has also contributed to D2.8 First Data Policies Recommendations.

The next period of work within the ELIXIR-CC will focus on:

- Demonstration of the RDSRS setup of EBI, CSC and CESNET to other ELIXIR Nodes and to EOSC-hub members, support them achieve similar setups and together with them define roadmap for improving the system capabilities and the integration with EOSC-hub common services.
- Deciding about virtual access and cost models between ELIXIR and EOSC.

- With the EOSC-Life project starting (March 2019), future interactions between EOSC-Life and EOSC-Hub will also be aligned via the ELIXIR-CC. Supporting the EOSC-Life Science Demonstrators with the ELIXIR RDSDS is one possible interaction. Decision on support will be taken after the technical requirements analysis of EOSC-Life demonstrators are finalised (Q3 2019).

1.2 Fusion

1.2.1 Ambition

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, 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.

1.2.2 User stories

No.	User stories - Fusion
US1	The Fusion CC wishes to demonstrate making use of EOSC computational and storage resources for running containerised modelling applications (primarily HPC and HTC). This requirement derives from the fact that local resources are not scaled for peak demand and we wish to use the infrastructure provides by EOSC (and public cloud providers) as a scalable, non-vendor specific resource. At a high level, this is an opportunistic use case where we wish to make use of any spare resources at sites, and thus going through an ordering process would be non-optimal since the user would not know local resources are exhausted until they submitted their job. It maybe that some sort of framework agreement would be needed between the community and the sites to allow this opportunistic use beyond the small number of cores already presented through EGI. Different parts of the workflows may involve different computational requirements, from simple single core machines to many core/multi-nodes. In the first case we would request resources on a single site, but only instantiate the number of machines required for a specific element of the workflow. It is desirable to develop this further to allow the instantiation of machines for different parts of the workflow to meet the requirements of each step. We are also interested in using both using traditional workflows and workflows within the ITER Integrated Modelling and Analysis Suite (IMAS) which is anticipated to become the standard framework for both modelling and analysis work in the future. In most cases the steps of the workflow communicate through files, with each stage producing its own unique file which acts as an input to the next stage. While running at a single site, it would be possible to request storage at that site of appropriate size. However, in the desirable case of different stages of the workflow running at different sites, either storage system accessible to many sites will be required for these intermediate files, or they would need to be transferred between sites. Final output data (and possibly intermediate data) should be accessible to the end user.

US2	As a site data manager, I am looking at how to improve access to experimental data for my users, and for other stakeholders even beyond the fusion community. However, in common with most science disciplines my site places an embargo on experimental data to allow researchers to publish. In addition, some data will not be made public where it has no scientific value (engineering tests for example), or where work is done on behalf of industry. Significant analysis work is performed on the MARCONI/Fusion supercomputer based at CINECA and for data sets which will be accessible it would be beneficial to my users if data could be hosted there. In addition, we want to offload public data to partner sites for hosting and access to the wider science community and general public, so that data used by the fusion community is kept on site and only accessed by fusion users. This combined with the restricted roadmap for tape technologies is pushing me towards replication as a means of bit preservation in the longer term. the community already has a data access mechanism (UDA) which it uses and must be usable at each site where the community will access data. General access will be via HTTP. Thus, as a site manager I would ideally like to put a full copy of my data on a trusted site which will prevent unauthorised access to the data (although not the metadata) during the embargo period but will make it accessible following that. That site should be able to provide me with data download statistics on an annual basis as part of my reporting to senior management and fundholders. Additionally, I would like that data to be copied from the trusted site to CINECA so it can be used optimally for analysis on the MARCONI computer and I would like a third copy to ensure there are four copies on my data availability for high availability (one local and three off site copies).

1.2.3 EOSC-hub services assessed, integrated

EGI DataHub; B2SAFE; B2DROP; DODAS; INDICO-PaaS; B2DROP

1.2.4 External services assessed, integrated

Singularity containers; Dynafed

1.2.5 Progress, status, plans

During this period the Fusion Competence Centre have made assessments of various technologies provided by the EOSC for potential use within both existing use cases and projected use cases in the ITER era. In the first period we assessed the following services and technologies:

- Cloud access with DODAS Thematic Service: initially seemed quite relevant to existing use cases. However, based on a paper-based exercise it was found that it did not provide the required functionality as it was limited to a single cloud site, while out use case involved dynamic selection based on data locality. However, the DODAS service may be relevant in the future.
- Storage access with OneData: With the last release, OneData has been installed successfully with replication between CEA and PSNC. Initial tests showed an unacceptably low write

performance, but this was tracked to a firewall issue at CEA. This has now been corrected and tests have demonstrated.

- Storage access with B2SAFE: Testing of the technology continues to prove difficult due to the complexity of interconnecting the three preferred sites. While this is ongoing, we have recently installed three containerised B2SAFE instances and a Handle server on a single cloud so that we can carry out functional tests. The B2SAFE instances have been configured as an iRODS federation and we have successfully tested replication between the different instances. Other tests from our data management test plan are ongoing. From the community's perspective, some essential functionality is missing from the basic service and would require additional community development, particularly 'self-healing' if a particular replicated object became corrupted (or deleted) either randomly, accidentally or through external bad actors. The additional effort to implement this has not been costed within this competency centre, and as such is currently not being pursued.

On the computing part we now have containerised versions of the ITER Modelling and Analysis System (IMAS) which is being adopted as a standard amongst European and many international fusion sites and are integrating code into this ready for deployment on EOSC-hub computational resources. On the non-IMAS workflows, we are working on the containerisation of two applications; one is a production workflow to from the MAST tokomak converting raw data to physical parameters, and the other is using a workflow called JINTRAC[1] which provides coupled simulations of different plasma scenarios within a tokamak. The first presents a challenge for running on cloud environments due to the use of licensed software and part of the work is replacing those components with open source alternatives. The second presents a challenge as JINTRAC is an extremely complex suite of coupled codes. We have already in this period looked at various container technologies using a component of JINTRAC using 8 processors.

During the next period, we will complete the testing of the storage systems, working with B2SAFE partners to achieve interoperability and completing formal testing on OneData. We will also continue to work on the implementation and testing of containerised workflows in EOSC resources, as mentioned previously. In addition, we will investigate the use of Dynafed for federated access to data in both OneData and B2SAFE which will minimise the need for data movement. The real challenge from the computational side is expected to be around accessing data; however, some of the partners are also involved in another project (FAIR4Fusion) which aims at making data more accessible not only across the community itself, but also to external organisations.

1.3 Marine

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…).

[1] https://doi.org/10.1585/pfr.9.3403023

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:

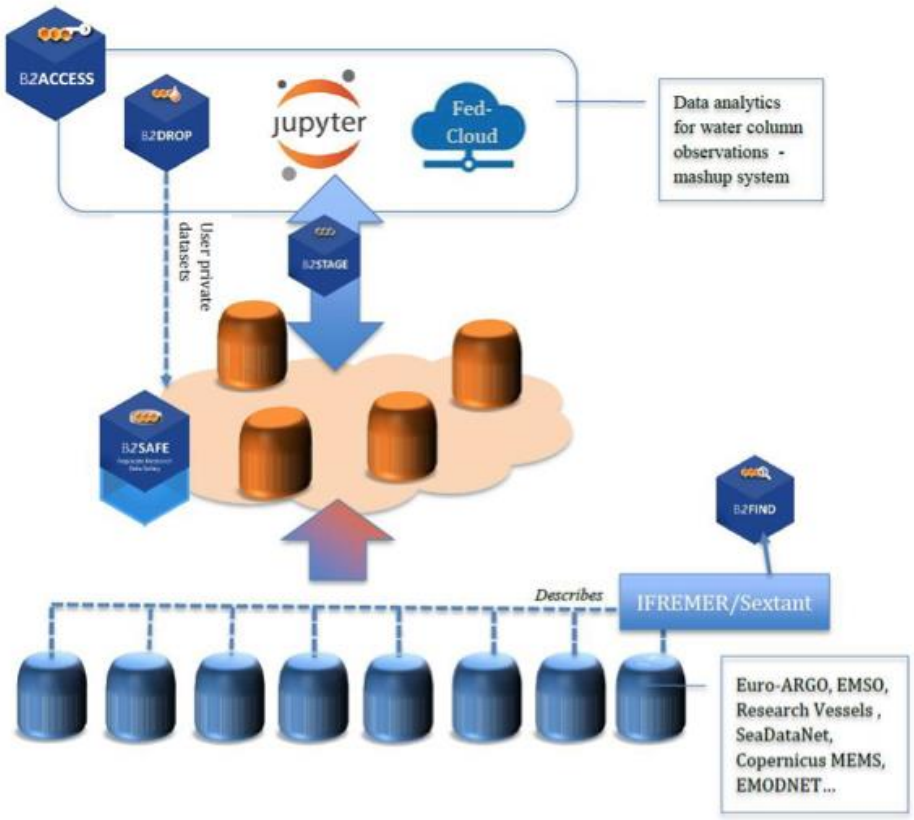
1. Making Argo data more easily accessible for subsetting and online processing. IFREMER and its partners work on this area.
2. Simplifying/harmonising the access to data that reside at SeaDataNet partners from cloud-based applications. MARIS and its partners work on this area.

1.3.1 User stories - Euro-Argo

The Marine community produces diverse types of data (typically time-series data). They wish to store those data in files and make these files easily browsable and accessible by researchers. To maximise ease of use the files should be made available to users via a Dropbox-like system that makes relevant data files visible for each user in his/her 'personal folder'. The users should be able to define patterns that define what kind of data they are interested in (location, time period, provider network, etc.) and the system should perform pattern matching to decide whether or not to make a particular incoming file (or set of files) visible for a given user. Such pattern matching can be CPU-intensive when we scale up too many users, many files files with complex data records. Depending on the community the source of data can be a single instrument (site) or can be multiple collection/production sites. In the latter case the data originating from multiple locations should be brought onto common formats and must be described with metadata in a coherent fashion.

The Argo activity of the Marine CC is testing (See Figure below)

- a combination of B2Find, B2Safe and B2Stage for the data management part (storage and transfer)
- a Jupyter, B2Access, EGI Cloud combination for user exposure. (data subscription and access)



No.	User stories - Argo
US1	A data provider should be able to link its data production instruments into the 'back-end' of the Marine CC setup and become a data provider for the CC users.
US2	A scientist should be able to browse the connected data source networks (e.g. Argo, EMSO, SeaDataNet, etc.) and define preferences for the data records he/she is interested in. The system should make matching records visible in his/her personal access folder.
US3	A user should be able to access his/her personal data access folder via a Jupyter system and perform data analytics on the data.

1.3.2 User stories – SeaDataNet

The current workflows of the SeaDataNet are based on a pre-cloud architecture. Many operations happen asynchronously and in batch mode. In order to better serve the Marine community, we want to provide fast and scalable access to the datasets. To improve the current workflow users should be able to take advantage of the improved access and availability of the cloud. A user should be able to store their data on a Dropbox-like environment, However, still, be able to process and

analyse them using both legacy/desktop software created during previous SeaDataNet projects and new cloud-based computing services. Furthermore, users should be able to discover data relevant to their needs using (semi) real-time discovery tools. Instead of preparing datasets for download for each user request, Data providers should be able to have their data stored on the cloud and provide access to users that have been granted permission. A data provider should be able to fix partial errors within a dataset during import, without having to re-upload the complete dataset.

No.	User stories - SeaDataNet
SeaDataNet user stories	
US4	As a user, I want to be able to use my legacy/desktop software to process and analyse data stored on the cloud.
US5	As a data provider, I want to only have to update erroneous files during import to only transfer the data required once.
US6	As a user, I want to be able to access my requested data through cloud computing tools within my cloud environment.
US7	As a user, I want to be able to find relevant datasets available within the cloud environment in (semi) real-time.

1.3.3 EOSC-hub services assessed, integrated

Argo: Jupyter (EGI Notebooks?), B2SAFE, B2DROP, B2Find, B2ACCESS

SeaDataNet: EGI DataHub

1.3.4 External services assessed, integrated

Argo: Cassandra, Elasticsearch

1.3.5 Progress, status, plans

EOSC-hub Marine Competence Centre (MCC) works on the design, integration, and dissemination of new, community-specific service platforms for the Argo and SeaDataNet communities. In period 1 the Argo team assessed ‘common services’ of EOSC-hub and external 3rd party and for their relevance for the Argo Data Discovery Platform:

- Euro-Argo data are synchronized continuously from Ifremer to B2SAFE: Argo data are daily transferred to CSC B2SAFE via its HTTP API. Data area compressed on the fly by the transfer protocol, the transfer “block” is a file containing one month of data (2.5GB) and the average

transfer time is 15 minutes, which is sufficient. Every day, the last two months are updated (5GB). Once a month, the historical files that have been updated are pushed to B2SAFE.

- User personal dataset uploaded in B2DROP: No issue to mention, this is a basic service from CSC that is operational for MCC.
- High performance parallel access to read/filter data with Elasticsearch index and its "data discovery API", Cassandra database, its "data charts API" api and its "subsetting API", "GDAC discovery" web application. Elasticsearch contains Argo metadata (2 million vertical profiles). Cassandra contains Argo data (5 billion individual observations from the 2 million vertical profiles). Ifremer dockarised and tested the setup on its own hardware and is now ready to deploy this service on EOSC-hub e-infrastructure. CSC, IN2P3 and CINECA are considered (service providers in the CC).
- Jupyter on top of B2SAFE/B2DROP, data access API with Python or Julia (DIVA). (B2DROP as an output for data results.) First application to use the setup: analysis of the regularly published DIVA map of Argo oxygen data. A DIVA demonstrator was installed on CINECA Jupyter infrastructure. Alternatives are considered: EGI Notebooks service and Ifremer has recently setup a JupyterHUB VRE. (Within EU project Marinet, Ifremer, CSC and Maris are in charge of defining the E-infrastructure of the European test sites. The 2 major functions are to publish datasets with DOIs and push them on a JupyterHUB VRE.)

The SeaDataNet activities started in PY2 to investigate solutions to eliminate bottlenecks in the SeaDataNet distributed data workflows developed by MARIS. Four (4) use cases (bottlenecks) were identified, and OneData is tested as potential solution for these bottlenecks. Computing and storage resources are provided for the pilot by CINECA, INFN-CNAF, IN2P3 from the CC, and from CYFRONET as external contributor.

The Argo partners in the CC will focus on the following activities in the next period:

- Define specifications for the required Hadoop-Spark-Cassandra-Elasticsearch infrastructure for the e-infrastructure sites and deploying those at the CC partner sites of CSC, CINECA, IN2P3. Register the service in the EOSC Portal and open it for early adopters in EOSC.
- Outreach and community engagement: Promote the achievements (service demonstrators) and the lessons learnt to the Argo community.
- Launch the development activity of the Data subscription service GUI at Ifremer. (recruitment must conclude)
- Define the specification of the Jupyter installation that's required for the users to interact with the Argo data that's staged to the e-infrastructure sites. Complete the Jupyter setup (one central, or multiple instances - to be decided)
- Expand the data platform with the new capabilities and according to user feedback.

In addition to the above, the CC partners involved in the SeaDataNet-Onedata pilot (MARIS, CYFRONET, CINECA, INFN-CNAF and IN2P3) will focus on:

- Finalising and formalising the resource provision for the pilot
- Implement and test Onedata for the identified use cases
- Report on the outcomes of the piloting activities

- Outreach and community engagement: Promote the achievements and the lessons learnt to the SeaDataNet community.

1.4 EISCAT_3D

1.4.1 Ambition

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 by M18 of EOSC-hub project (from Sept 2019).

1.4.2 User stories

EISCAT_3D will generate raw data at up to three radar sites (each with almost 10000 antennas) and has to make those data browsable and analysable for researchers, as well as archiving the data for the long-term. It is planned to archive around 2 PB per year. This primary mission is achieved by procuring sufficiently big storage and network capacities at a few data/compute centres (for back-up or load distribution). The data will be transferred to this/these locations after initial filtering and calibration at the data source(s). The data/compute centres are responsible for data archival, data curation, generation of derived data (level 1-2-3) and for sharing the data with scientists.

Data must be shared with scientists via a data portal and programming APIs. The EISCAT_3D Data Portal should offer the following main features:

- AAI, user login
- Data browser
- Data download
- Online computing (analyse data without downloading them, using cloud resources, reference applications or user's own software)

The CC is exploring the use of the following services within the framework of EOSC-Hub:

- Application and data catalogue portal with compute integration: DIRAC
- Metadata catalogue for high-level data: B2Find
- Metadata catalogue for low-level data: DIRAC
- Compute resources: EISCAT, cPouta cloud at CSC
- Data storage and transfer (data → compute centres): File transfer TBD, e.g. B2STAGE
- User SSO: EGI CheckIn and B2Access

No.	User stories
US1	Any researcher should be able to access the portal and browse metadata. The portal grants/denies access to data and processing based on affiliation. (Meta- data should be available for all researchers. The real data for authorised users only.)

US2	Authorised researchers should be able to select the EISCAT_3D data they are interested in for download or for analysis.
US3	Authorised researchers should be able to browse reference applications in the portal, select an application for use, feed their data in (from US2), visualise or download the analysis result.

1.4.3 EOSC-hub services assessed, integrated

Cloud (EGI or CSC), EGI Workload Manager, EGI HTC, EGI Data Transfer (possibly with Rucio), B2SHARE, EGI Check-in

1.4.4 External services assessed, integrated

Containers, dCache

1.4.5 Progress, status, plans

This task aims at developing the EISCAT_3D portal based on the DIRAC interware to enable users browse and analyse EISCAT_3D data. Given that EISCAT_3D data will not be available before the system is operational around 2021, the DIRAC storage that is used in the CC provides access to existing EISCAT data. The DIRAC GUI web portal was adapted for EISCAT by the EISCAT_3D CC of the EGI-Engage project. This development continues in the EOSC-Hub CC. During the first period the work concentrated on three topics:

- Integrating the DIRAC web portal with the EGI Check-in service, in order to remove the personal X.509 certificate requirement for user access. A working prototype of the integration was achieved in Q1 end 2019 and will be put into production in the summer of 2019.
- Arranging IaaS cloud resources at CSC (a CC member), in order to deploy compute intensive data analysis algorithms at CSC using the DIRAC work management system. This has been demonstrated with the existing graphical animation software (real-time graph) for spectral EISCAT data.
- Defining a metadata architecture which can store and link multiple levels of metadata descriptions about EISCAT_3D data, even if the different data levels are physically stored in different systems. The basic metadata sources will be the local databases at the EISCAT_3D sites. Semantic mapping to standardized metadata formats and a persistent identifier (PID) scheme will be developed following recommendations from ENVRI-Fair. The DIRAC file catalogue can accommodate user-defined metadata. It is also planned to make the data searchable in B2FIND.

The team works with the EISCAT_3D low level software developers as well as with partners in the Nordic NREN collaboration (Nordunet), in order to test scalable storage systems (dCache etc), define data formats and benchmark data transfer.

The upcoming tasks of the EISCAT_3D Competence Centre are:

- Roll out the federated AAI integration into the production version, register the service in EOSC Portal and EOSC-hub Marketplace, open it up for user access in WP13.
- Deploy data analysis routines on computing resources at CSC and connect these to the DIRAC portal (incl. Single Sign-on). For this prototyping we will package existing software, in addition to data plotting also for instance lag profiling software, to allow job submission and data transformation rule setup with DIRAC.
- Further elaborate the metadata model and its implementation in a hierarchical DIRAC - B2Find metadata catalogue system.

1.5 EPOS-ORFEUS

1.5.1 Ambition

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.

1.5.2 User stories

No.	User stories - EPOS-ORFEUS
US1	As a **provider** of an EIDA data centre I want to provide users with an authentication and authorisation service in order to enable them to securely access restricted and embargoed data.
US2	As a seismological **researcher** I want to search for datasets offered by EIDA and stage them on the available cloud infrastructure offered by EOSC providers.
US3	As a seismological **researcher** I want to analyse my data in a Jupyter environment, pre-populated with my preferred libraries and with access to my pre-staged datasets. I want to store results in my personal workspace/storage area and eventually share them with my colleagues.
US4	As an EIDA **data manager** I want to define my data management (DM) policies and share them with my colleagues at EIDA data centres. I want to enable them to understand, adjust and apply DM policies at their data centres.

1.5.3 EOSC-hub services assessed, integrated

Jupyter (EGI Notebooks and custom), B2SAFE, B2STAGE, B2HANDLE, B2DROP, B2ACCESS

1.5.4 External services assessed, integrated

Discovery (PIDs)

1.5.5 Progress, status, plans

The activities carried out by the EPOS-ORFEUS-CC during the first reporting period of the project targeted four main areas that were identified in the initial work plan and are described below.

- Federated AAI (FedAAI) – Investigations were performed for the establishment of a federated AAI system. A component from the EOSC-hub service catalogue was chosen, i.e. B2ACCESS, and integrated with seismological data services (4 types of webservices to provide standardised and open access to data). The resulting technical solution was first piloted by engaging a selected target user community (AlpArray project users), and then rolled-out as an operational service.
- Data Staging service (DaSta) – Scenarios and use cases were defined; technological selection was performed, and a prototype was developed. This exploits a community metadata catalogue (i.e. WFCatalog) for discovery, persistent identifiers (PIDs) issued via B2Handle, B2STAGE and replicated seismological archives with B2SAFE. Preliminary tests were performed. These showed feasibility of the developed solution which can fulfil the identified requirements. Some issues remain to be solved at e-infrastructure level (such as communication between staging nodes; allocating staging nodes for iRODS) in order to achieve a stable and fully working distributed deployment.
- Enabling scientific data analysis (UGeP) – The requirements for the generation of user-defined products (close to data location) were defined and technological solutions were identified and selected. Two different solutions based on Jupyter Notebook were tested: EGI Notebook deployed at GRNET and ResearchDrive with Jupyter integration at SURFsara. By running selected applications, several issues arises that need to be addressed by e-infrastructure providers. These include harmonisation and integration of analysis environments, better integration with data sources, user management and improving robustness and stability to cope with scale.
- Federated Data Lifecycle service (FeDaLi) – A framework, named RuleManager, was developed that enables data centre operators to define and run data management policies. Such policies can be represented and shared as JSON documents. The framework can integrate with B2SAFE and broadly used seismological components and services. This has been tested with a selection of policies.

Additional activities include (1) the definition of an architectural plan for the integration of the services: UGeP, DaSta and FeAAI, (2) the setup and testing of a new B2SAFE installation between NOA and GRNET and (3) the update of the installation between KNMI and SURFsara.

The plan of the EPOS-ORFEUS-CC for the next period includes the following tasks:

- FedAAI – Monitor the usage of the developed solution, offer support and plan integration with EPOS ICS AAI.
- DaSta – Finalise the integration of the different components, solve issues and test the technical solution with a broader target user group.
- UGeP – Address issues and work to provide a working platform to be evaluated by selected users with different analysis software.
- FeDaLi – Extend the evaluation to all the data centres of the CC and work on the publication and sharing of the policies.

1.6 Radioastronomy

1.6.1 Ambition

The Radio Astronomy CC will 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.

1.6.2 User stories

No.	User stories - Radioastronomy
US1	As an Observatory, we want to offer Single Sign On to our users and manage community access through a central federated collaboration management service, such as COmanage, to improve user experience and consolidate user administration for services.

US2	As a scientific user, I want to perform LOFAR data analysis on archived data-products using available scalable compute infrastructure, allowing for long-term storage and inspection of results. It must be possible to automate initiation and monitoring of processing workflows including data staging and storage. I want to use portable software deployment such that time spent on porting applications is minimized, an integrated workflow management framework, and user workspace to store data products that are to be evaluated or further processed.
US3	As a scientific user, I want to enter science-grade data products in a science data repository that supports the FAIR principles to ensure long-term data preservation and attribution of effort. This will further improve sharing of data with colleagues and access to data from other science domains. It should be possible to access data in the science data repository using direct links to individual data objects via an anonymously accessible public URL such that other services, e.g. those provided by the Virtual Observatory, can be built to provide access to the data.

1.6.3 EOSC-hub services assessed, integrated

B2SHARE, B2FIND, Check-in

1.6.4 External services assessed, integrated

dCache, Common Workflow Language with Singularity containers

1.6.5 Progress, status, plans

A high-level target architecture has been defined for user-oriented services for the LOFAR community, considering the main service components that are to be integrated and further developed in the context of the EOSC-hub project. Proof of concept development work has been carried out to evaluate applicability of service components to the LOFAR domain, and as the outcome the following baseline solutions have been identified for the various technical areas:

- AAI: LOFAR services will initially build on the EGI Check-In AAI services, using its COmanage to support administration and self-organisation of collaborative projects.
- Storage: Storage services will build as much as possible on dCache functionality, dCache being the storage middleware adopted by all LOFAR Long Term Archive infrastructure providers.
- Repository: A science repository solution based on EUDAT service (B2SHARE and B2HANDLE) will be set up and investigated in a pilot with a selected number of researchers. EPIC has been chosen as the baseline candidate for a Persistent Identifier service.
- Processing: The most appropriate approach to implementing processing workflows is still under investigation. In the EOSCpilot, a solution based on applications deployed via (Singularity) containers and workflow definitions in the Common Workflow Language has

been demonstrated to work at functional level but still is to be demonstrated at relevant LOFAR scale.

A pilot AAI integration with EGI Check-In has been implemented and demonstrated for web-based services. Particular attention has been given to the data staging and retrieval services which allow community members to request data from the archive to be staged from tape medium to disk in order to retrieve it for inspection or further analysis. Data retrieval requires users to use a custom HTTP based data retrieval server, requiring the custom LOFAR user account, or use personal X509 'grid' certificates in conjunction with GridFTP tools. The latter is the more scalable and performant option but obtaining, and maintaining, X509 certificates is perceived by many users as cumbersome and using grid tools is not always trivial. Moreover, support for GridFTP is ending soon. A replacement service has been piloted that integrates with the federated AAI and supports user data retrieval based on authorization tokens in conjunction with a scalable/performant WEBDAV interface that current versions of dCache support. AAI integration with the Oracle-backed LOFAR archive catalogue has been found to be more complicated, in particular as it builds on the Oracle user administration for authentication and authorization which is provisioned through a local Identity Management system and is still under investigation. The applicability of B2SHARE and B2FIND to offer science data repository services has been discussed with SURFsara and the service maintainers. The current plan is to implement a demonstrator that offers a B2SHARE/B2FIND interface for registration and discovery of data generated by/for the LOFAR community, and existing dCache systems for the underlying storage infrastructure. Maintaining constant links between dCache and the PIDs used in B2SHARE/B2FIND will require customization of existing functionality in these services.

The CC will focus on the following tasks in the next period within the various technical areas:

- AAI: Integration of operational services with Check-in. Take the new stager service, providing token-based access to LOFAR data via the dCache WEBDAV interface, into production and identify/implement a method to integrate the Oracle-based catalogue with EGI Check-in.
- Storage: Offer token-based access to dCache services as a user service and setup user/data repository workspaces within dCache.
- Repository: Implement pilot instances supporting the LOFAR community with a B2SHARE science repository and an EPIC Persistent Identifier service.
- Processing: The evaluation of portable workflow frameworks will continue with the objective to decide on an implementation to be supported for the LOFAR community in Q4 of 2019. As a result of project interdependency and availability of personnel, demonstration of a CWL/Singularity approach at LOFAR production scale has been postponed and is now planned to be evaluated in Q3 2019. Alternative approaches that are being investigated in concurrent projects include the custom LOFAR ‘Generic Pipeline Framework’ that is currently being used to process one of the largest LOFAR science cases but is at the moment not supported, data processing frameworks under evaluation for the Square Kilometer Array within the AENEAS project.

1.7 ICOS-eLTER

1.7.1 Ambition

The ICOS - eLTER Competence Centre groups two scientific communities:

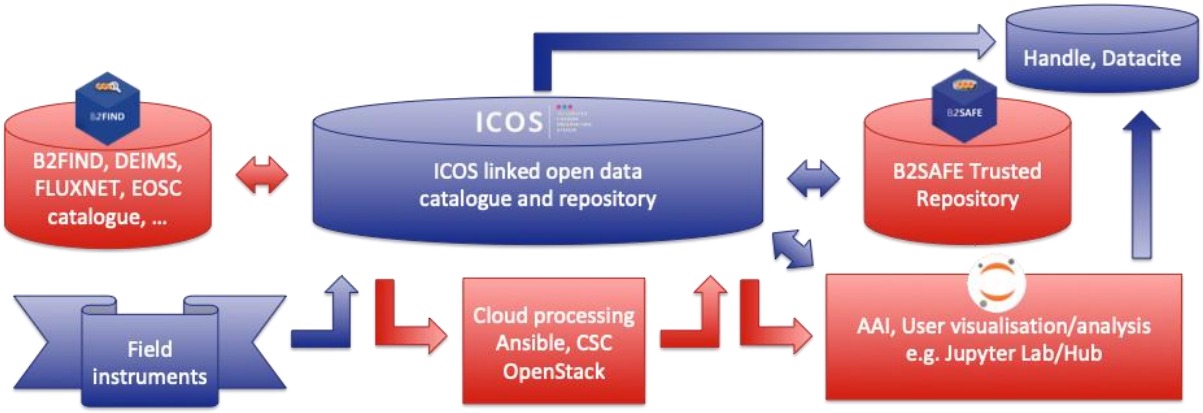
1. The Integrated Carbon Observation Systems (ICOS) research infrastructure[2]
2. Long-Term Ecosystem Research in Europe (eLTER) community[3]

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.) The resulting data is to be automatically ingested and be published in the ICOS Carbon Portal and at the Thematic Center. At this stage the integration of the data flow from EAA can also start. The same data workflow adapted for final data quality data will be implemented in February 2019. The final production ready version of the workflow should be finished by M18.

The eLTER group aims to develop and share data analysis 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.

1.7.2 User stories – ICOS

The ICOS Carbon Portal should enable researchers to access carbon data from the ICOS thematic centre field instruments, and to process and visualise those data using the portal GUI. The expected underlying technologies are represented on the diagram below.



[2] https://www.icos-ri.eu
[3] http://www.lter-europe.net/

1.7.3 EOSC-hub services assessed, integrated

ICOS: Cloud (CSC), B2SAE, B2STAGE, B2SHARE

eLTER: Cloud (EGI), Jupyter (EGI Notebooks), B2SHARE, B2FIND

1.7.4 External services assessed, integrated

eLTER: OGC SOS Data service (provided by eLTER Data Nodes)

1.7.5 Progress, status, plans

ICOS:

Continued with using the B2SAFE/B2STAGE at CSC using the improved http API to the data object storage workflow in parallel with the production level iRODs data storage at PDC. Data transfer to the B2SAFE implementation was quite reliable, although besides the previously reported interruption and a few announced maintenance periods, a halt in the service was discovered in May 2019. It turned out that the storage service had not been working for several weeks. After restart and another stop, the problem turned out to be limited memory of the server at CSC. After increasing the memory, the service worked again. Since this major interruption a few more hiccups have been detected, so the reliability of the B2STAGE/B2SAFE service still needs to be improved. The user case of the ICOS CC aims to use the CSC cPouta Cloud environment for processing the daily ecosystem flux data. The aim is to transfer the raw ecosystem flux data directly from the data loggers to the Carbon Portal into the data staging area. From the staging area the data is transferred in daily packages to the ICOS CC object data store that is connected to B2SAFE as its trusted repository. The processing of those data was developed in ENVRIplus based on the D4Science platform, and the orchestration tools used have been developed for the ICOS STILT footprint tool based on the EGI grid computing services. Personnel from the ICOS Ecosystem thematic centre have received training and are now working with the orchestration and integration with the ICOS semantic link framework that gives access to the input and output data through the B2SAFE repository. It is expected that the actual flow and processing of the data can be implemented in the second half of 2019. Initial tests performed at ICOS Carbon Portal own servers indicated that the processing of the daily raw data from 100 stations into the near real-time product could be completed within the hour, using only 20 cores. At this moment the launching of virtual machines at CSC CPouta has been tested, and the required software based on the OpenStack API has been developed without any major problems, indicating that running the data processing can be implemented as soon as the needed scripting of the data workflow for the processing software is ready. The use of cPouta virtual machines has also been tested by implementing other applications (LPJGUESS, a dynamic vegetation model) than the planned flux data processing.

The CC expects to run with actual near real-time data and full production or near real-time data in the second half of 2019. The resulting data is to be automatically ingested and be published in the ICOS Carbon Portal and at the Thematic Centre. At this stage the integration of the data flow from EAA can also start.

eLTER

Activities started in PY2 and focused on the implementation of the "Data Validation Lab" use case, i.e. ingesting online data sources from OGC SOS compliant services. (e.g. https://cdn.lter-europe.net/data/sensor?id=/ECN/T12/DRYTMP_RH/2&service=observations) The work involved formulation of the use case and the evaluation of the required infrastructure provided by EOSC-Hub. The use case requires access to time series data from a range of eLTER sites providing multiple sensor Services in near real time, as well as access to modelled and interpolated data from these eLTER sites. It is aimed to provide an environment to apply descriptive statistics methods for time series data and enable data quality flagging (analysis and QA workflows) as well as to identify anomalies in data (e.g. data errors, events, changes points) from multiple data sources applying a range of data analytics methods (e.g. univariate and multivariate statistics). The implementation of the use cases focuses on development of the workflows as well as on evaluating the technical barriers on migrating the service to a different cloud infrastructure and provider in the EOSC context. The workflow will be prototyped using EGI Notebooks (with R environment) infrastructure. The implementation of the use case workflows is foreseen for the second half of 2019. Migration to production at FZJ will be evaluated after the prototyping phase.

The group will focus on the following tasks within the next period:

- Finalisation of the data Validation workflows
- Implementation of the eLTER Data Validation Workflow using EGI Notebooks (in the second half of 2019)
- Testing of performance for data provision using data Services (e.g. OGC SOS) for different temporal and spatial resolutions
- Assessment of options to migrate "eLTER Data Validation Use Case" to production cloud Providers, particularly FZJ in Germany the site committed for eLTER (in 2020)

1.8 Disaster Mitigation Plus

1.8.1 Ambition

The competence centre brings together institutes from the Asia-Pacific region that work on modelling, re-modelling, predicting disaster events, with the ambition to be able to predict such events and to design mitigation actions against them.

1.8.2 User stories

The CC's user stories are basically resimulation of disaster events that happened in the Asia Pacific region:

No.	User stories
US1	Case study on **Sulawesi (Indonesia) tsunami happened on 28 Sep 2018**
US2	Case study on storm surge induced by **Tropical storm Pabuk (#36) in Thailand (3-5 Jan 2019)**
US3	Case study on long-distance **Dust transportation from biomass burning in Northern Thailand (2018)**

1.8.3 EOSC-hub services assessed, integrated

EGI Cloud, EGI HTC

1.8.4 Progress, status and plans

During the reporting period the CC focused on the scientific simulation, and on the technical infrastructure activity areas. The progress was:

- Case studies:
    - Tsunami: Sulawesi Tsunami (Sep. 2018) and tsunami risk estimation of countries around South China Sea are ongoing - prototype of Sulawesi case had been constructed. Validation with more observation data is to follow. Data collection and identification of target observation places around the South China Sea are in progress.
    - Storm surge: collecting observation data from few agencies, e.g., tidal gauge data. Events caused by tropical monsoon in 2006 or 2016 are candidate events for re-simulation and model validation.
    - Fire/Haze/Smoke monitoring and dust transportation: The biomass burning case of northern Thailand in 2017 was selected for re-simulation and model validation. Emission of PM10 will be traced in this study. We hope to be able to obtain more observation data to extend the analysis to CO, NOx, SO2, O-zone, PM2.5 and CO2, and hope to expand the simulation to upper Asia and Indonesia.
    - Flood: case of 2015 in northern Thailand is the first candidate for resimulation because some of observation data is already available at the partners.
- Technical infrastructure and simulation portals:
    - The CC continued to support the WRF-based weather simulation portal and the iCOMCOT-based tsunami wave propagation simulation portals. These use HTC computing resources from the Asia-Pacific VO of EGI and have been developed during the EGI-Engage project. The CC initiated the inclusion of these portals in the EOSC-hub service catalogue (through the WP2 SPM process).
    - The development of a new, storm surge simulation portal prototype concluded, and the setup is currently under testing. The new portal should be available in Q2 of 2019.

- Studying the interoperability of Asia-Pacific and European EOSC-hub services is underway and will continue in the next period. The goal is to define a roadmap for the adoption of the latest, EOSC-related European practices and interfaces in the Asia Pacific e-infrastructure community.

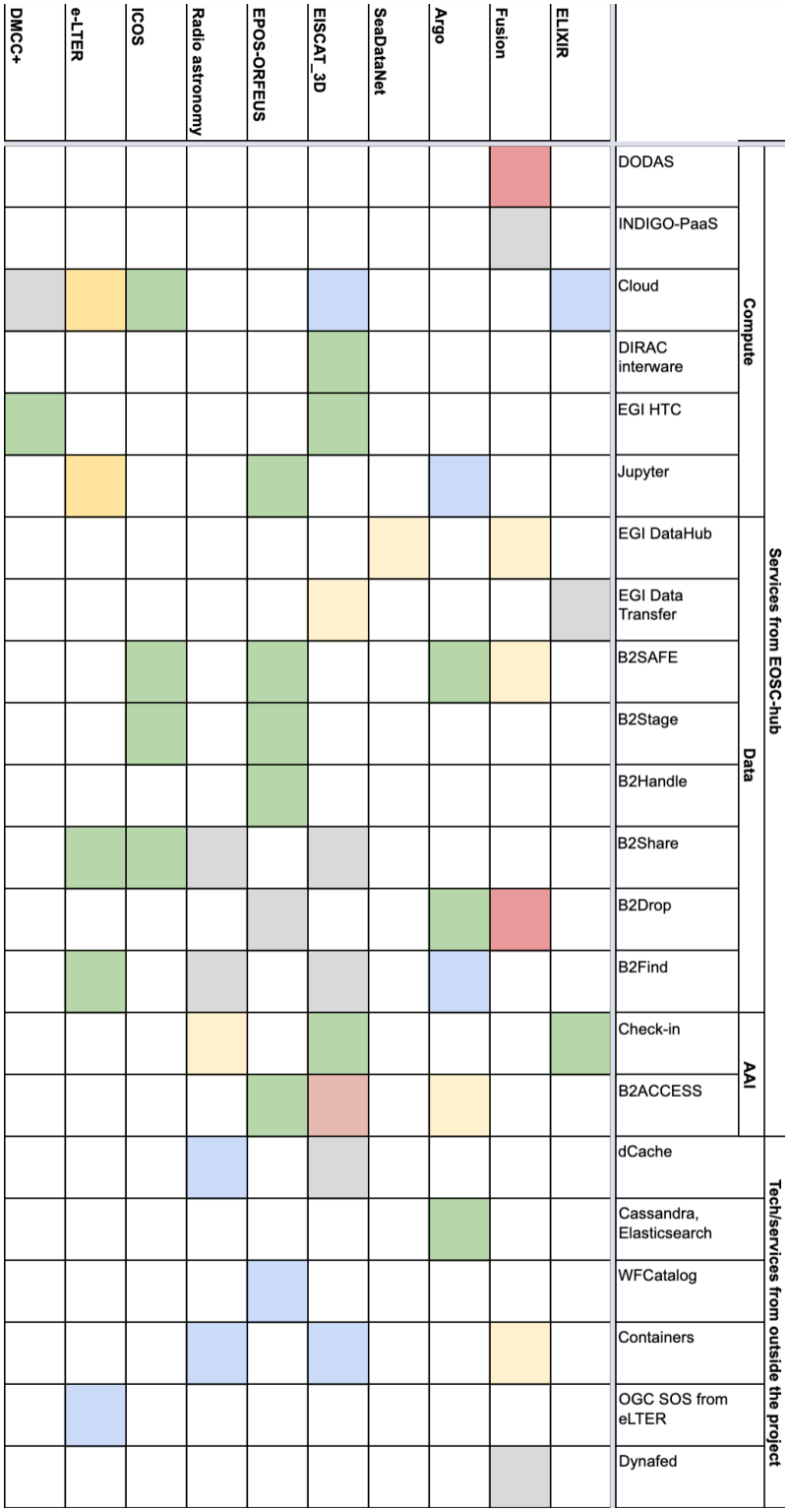
The CC organised an Environmental Computing Workshop and the DMCC+ face-to-face meeting in Taipei in colocation with ISGC 2019. The meeting helped the CC continue working on the two activity areas by performing:

- Scientific case studies: confirm the schedule of each case study according to the availability and quality of required data. Monthly meeting on each case study has been arranged.
- Technical infrastructure and simulation portals:
  - Conclude the registration of the weather simulation and tsunami simulation portals in the EOSC-hub catalogue.
  - Finish the testing of the storm surge simulation prototype portal (by the end March 2019.), then start adoption in the scientific case studies, and in the EOSC-hub service catalogue.
  - Additionally, to these the CC has the ambition to expand the collaboration with the Agriculture community in Asia and co-host a joint meeting at APAN47 between 19-21 Feb. 2019 in Korea.

2 Conclusions and observations

1. Despite there were hiring delays in a few of CCs (EISCAT_3D, Radioastronomy, ICOS-eLTER), there is only one CC (ICOS-eLTER) where this caused visible delay to the work performed. However, because this CC was originally envisaged to run only until month 18 of the project, the task (T8.7) can be extended until the end of the project, for another 18 months, without significant change in the description of action.
2. Most CCs are still in the middle of service evaluation, with only a few service evaluations results available so far. (See GREEN and RED cells in Appendix 2 for these evaluation results) Decisions about the suitability and adaptability of EOSC-hub common services is still to come in the second project period.
3. Promotion and training of new users, primarily power and test users, is still to happen in the second period (and in WP11).
4. Although at the start of the project most CCs foresee the integrated, joint use of services from EGI, EUDAT and INDIGO-DataCloud, such integrated use has not happened until now. The reason of this is, on one hand the lack of visible progress with enabling compatibility among EGI and EUDAT services by WP6 (e.g. AAI incompatibility), and on the other hand the possible service alternatives that EGI and EUDAT offer for the same functions allow CCs to combine services from a single provider. (e.g. DataHub and B2SHARE for data access; EGI Cloud and CSC cloud for IaaS; Check-in and B2ACCESS for access management).
5. Arrangements for long-term service provisioning is still to be arranged in all, but the ICOS CC. ICOS-ERIC ran a tender in 2018 and selected EUDAT Ltd for the provisioning of IT services (compute and storage) for the long-term. We yet to see whether and how WP12 (Procurement) and OCRE can support the CCs and their respective RIs in this endeavour.

Appendix I. Service adoption within the CCs



Colour codes:

***GREY**: technology/service is considered for adoption, but the integration and assessment work is yet to start*

***YELLOW**: integration is ongoing*

***RED**: technology was assessed and was found unsuitable*

***BLUE**: technology is integrated but user assessment is yet to finish*

***GREEN**: technology is integrated and positively evaluated by users*