EGI-Engage

Cross-Infrastructure case studies report

D4.8

Date	03 March 2017
Activity	WP4
Lead Partner	CSIC
Document Status	FINAL
Document Link	https://documents.egi.eu/document/3026

Abstract

This deliverable describes the EGI-EUDAT integration activities with a comprehensive report on all cross-infrastructure case studies undertaken following the call in PM15 of the EGI-Engage project. The deliverable will include initial integration work done to demonstrate basic interoperability from a technical point of view. We then cover the work done with the primary use cases, the Integrated Carbon Observation System (ICOS) and the European Plate Observing System (EPOS). Finally, we describe the activities to achieve AAI integration on both infrastructures.

The EGI-Engage project is co-funded by the European Union (EU) Horizon 2020 program under Grant number 654142 http://go.egi.eu/eng

COPYRIGHT NOTICE

DELIVERY SLIP

	Name	***Partner/Activity***	***Date***
From:	Enol Fernandez	EGI Foundation/WP4	
Moderated by:	Malgorzata Krakowian	EGI Foundation/WP1	
Reviewed by	Jens Jensen Diego Scardaci Alvaro López García	STFC INFN/WP3 IFCA/WP4	24/02/2017 24/02/2017 21/02/2017
Approved by:	AMB and PMB		3/03/2017

DOCUMENT LOG

Issue	***Date***	***Comment***	***Author/Partner***
v.1	11/01/2017	Initial TOC	E. Fernandez/EGI.eu
v.2	14/02/2017	First version with all the use cases described	M. Viljoen/EGI.eu M. Hellström/ICOS A. Spinuso/KNMI E. Fernandez/EGI.eu
v.3	1/03/2017	Version addressing external review comments	M. Viljoen/EGI.eu M. Hellström/ICOS A. Spinuso/KNMI E. Fernandez/EGI.eu
v.4	3/03/2017	Add information on data volumes handled by use cases	E. Fernandez/EGI.eu
FINAL	3/03/2017	Final version after PMB review	

TERMINOLOGY

A complete project glossary and acronyms are provided at the following pages:

- https://wiki.egi.eu/wiki/Glossary
- https://wiki.egi.eu/wiki/Acronyms

Contents

Executive summary

This deliverable reports the EGI-EUDAT integration activities with a comprehensive report on the cross-infrastructure case studies undertaken following the call in PM15.

EUDAT infrastructure provides research data services, training and consultancy for researchers, research communities, research infrastructures and data centres. Task 4.3 task seeks the collaboration with the service providers of EUDAT towards a harmonisation of the two infrastructures, including technical interoperability, authentication, authorisation and identity management, policy and operations. The definition of the roadmap for the collaboration is guided by a set of relevant communities who are already collaborating with both infrastructures infrastructure in the field of Earth Science (EPOS and ICOS). These communities were selected after a call for cross-infrastructure case-studies in PM15.

The first outcome of this activity has been the definition of a universal use case that covers the user needs with respect to the integration of the two infrastructures previously identified. The ICOS and EPOS use cases are still under development but both are already able to exploit EGI and EUDAT services for delivering services to their final users.

The integration of AAI mechanism of both e-infrastructures is feasible by using X.509 certificates, but EGI and EUDAT are migrating their AAI system to use federated authentication mechanisms and a plan for the integration of those has been defined.

1 Introduction

EUDAT[1] is a collaborative Pan-European infrastructure providing research data services, training and consultancy for researchers, research communities, research infrastructures and data centres. EUDAT's vision is to enable European researchers and practitioners from any research discipline to preserve, find, access, and process data in a trusted environment, as part of a Collaborative Data Infrastructure (CDI) conceived as a network of collaborating, cooperating centres, combining the richness of numerous community-specific data repositories with the permanence and persistence of some of Europe's largest scientific data centres.

The EGI-EUDAT collaboration started in March 2015 with the main goal to harmonise the two infrastructures, including technical interoperability, authentication, authorisation and identity management, policy and operations. The main objective of this collaboration is to provide end-users with a seamless access to an integrated infrastructure offering both EGI and EUDAT services and, then, pairing data and high-throughput computing resources together.

A call for cross-infrastructure case studies call was created by PM15 to collect interest on the integration of EGI and EUDAT services from relevant communities. The text was sent to EUDAT for comments and EGI and EUDAT selected a set of relevant user communities who are already collaborating with both infrastructures. These user communities are able to bring requirements and help to assign the right priorities to each of them. In this way, the integration activity has been driven by the end users from the start. The identified user communities are relevant European Research infrastructure in the field of Earth Science (EPOS and ICOS).

[1] www.eudat.eu

2 Cross Infrastructure Case Studies

2.1 Generic use case

The foundations of the EGI/EUDAT interoperability work were carried out in late 2015 and were termed the Universal Use case. This consisted of demonstrating basic interoperability between EGI and EUDAT infrastructures at the most fundamental level, predicated on the fact that both make use of the same underlying technologies (X509 digital certificates for authentication and GridFTP for data transfer). The Universal Use case was successfully demonstrated at the EGI conference in Bari by Diego Scardaci and paved the way for more complex use cases to be developed to address the real-life needs of the user communities.

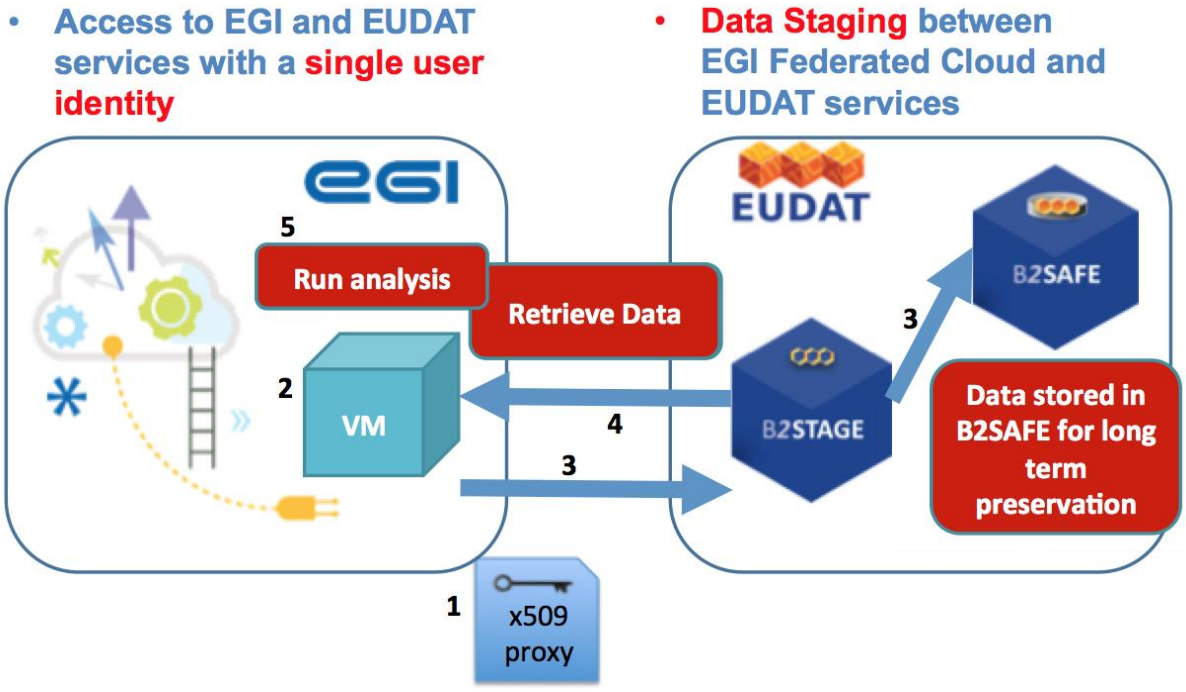

Figure 1 Universal use case

First an X509 proxy certificate is generated at the command line (1). Since both EGI and EUDAT infrastructure makes use of X509 digital certificates for authentication of end users and trust anchors are installed from the International Grid Trust Federation (IGTF) this step ensures interoperability at the authentication layer. A VM with a gsiftp client is then instantiated on the EGI Federated Cloud by issuing an OCCI command (2) and the proxy certificate copied into it. A sample file is then copied from the VM via EUDAT B2STAGE to the EUDAT B2SAFE service for long term preservation using the gsiftp protocol (3). The same file is then retrieved back using the same means (4) and compared with the original file to ensure that they are identical (5).

Although undoubtedly a very simplistic example, the Universal Use Case serves to demonstrate the essential compatibility with the EGI Federated Cloud and EUDAT data services.

2.2 ICOS

ICOS (Integrated Carbon Observation System[2]) is a European research infrastructure (RI) which aims to facilitate research to help understand the greenhouse gas (GHG) budgets and perturbations in Europe and adjacent regions. The overall data flow in the RI is shown in the Figure below. ICOS is based on the collection of high-quality observational data by measurement stations operated long-term (15+ years) as national networks in the RI member states. Besides the observational data products, which are distributed via the ICOS Carbon Portal (CP), also various "elaborated data products", i.e. outputs of modelling activities based on ICOS observations, are compiled and distributed by the CP. Furthermore, the CP facilitates the creation of such elaborated products by the research community. An example is the "footprint tool", which is an interactive, on-demand service that computes and visualizes the sensitivity of Green House Gases (GHG) concentration signals at potential and existing ICOS atmospheric measurement stations to GHG emissions and fluxes from different sources. This tool has been chosen as the initial use case in the collaboration with EUDAT and EGI.

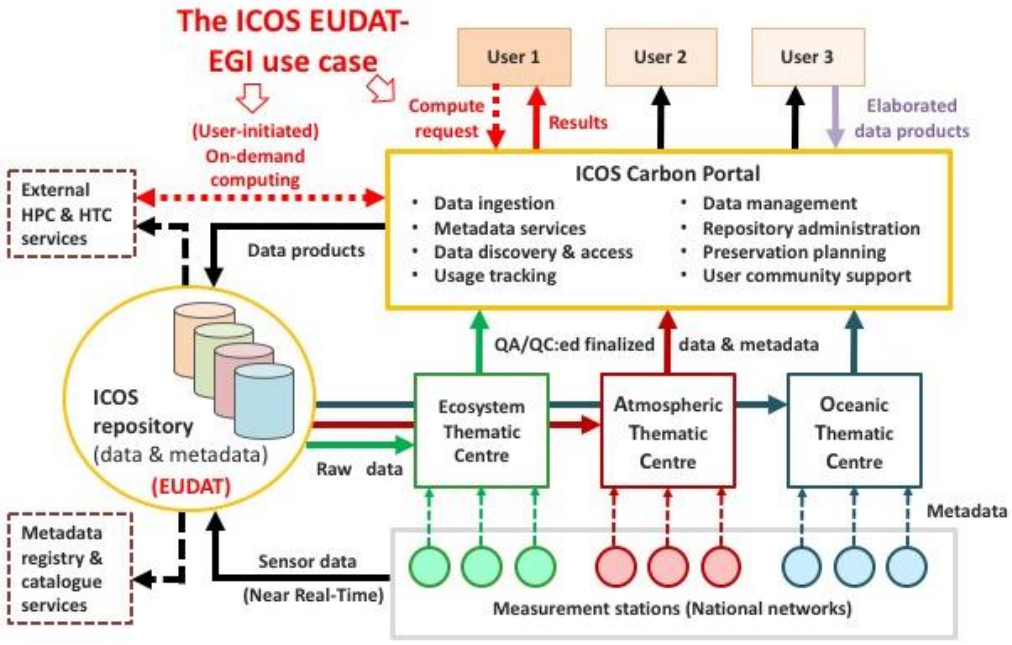

Figure 2 ICOS use case

[2] https://icos-ri.eu

2.2.1 Architecture

The central part of this service run by ICOS CP consists of a dockerized version of the atmospheric transport model STILT (Stochastic Time-Inverted Lagrangian Transport) and a web interface for the communication with the users. Both are run in Virtual Machines (VM) in the EGI Federated Cloud. The work and data flow in this use case is illustrated in Figure 3..

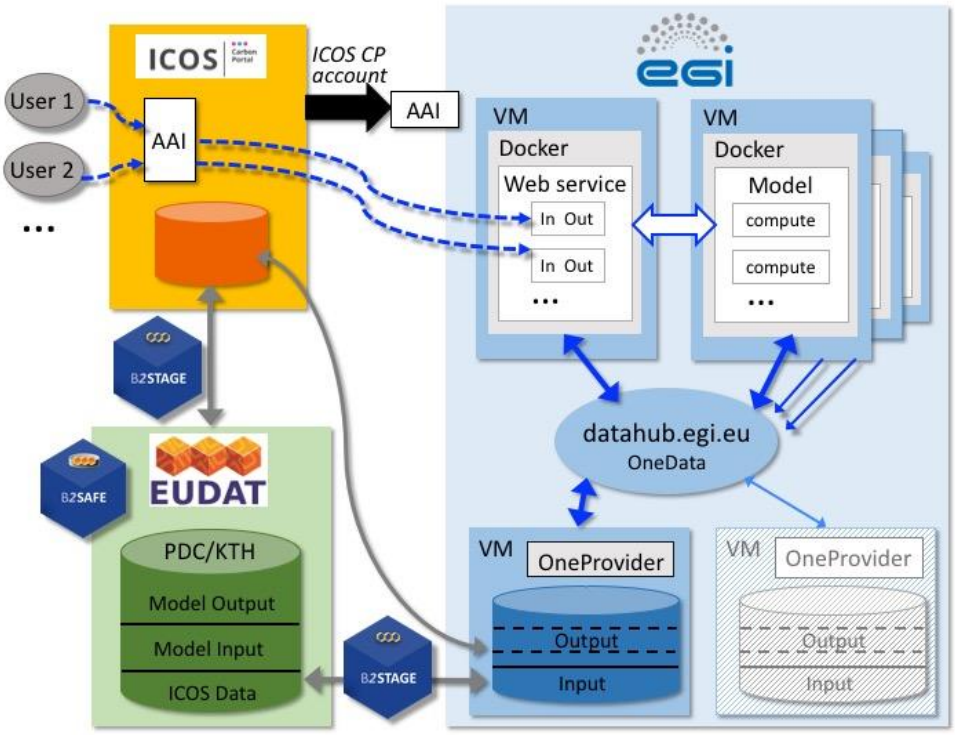

Figure 3 ICOS Carbon Portal Architecture

Users of the footprint tool access the service via the ICOS CP website, where also authentication and authorization is handled by ICOS CP. All interactions with EGI and EUDAT services are handled by ICOS CP on behalf of the user. Depending on the number of users and the workload on the VM(s) running STILT, more VMs for the model run might be instantiated and model jobs are distributed on these VMs.

The STILT model runs require input data form several data streams, which are first collected and pre-processed at ICOS CP and then stored for long term preservation in an EUDAT B2SAFE instance (at PDC Center for High Performance Computing at the KTH Royal Institute of Technology, Sweden). Data transfer is handled using the EUDAT B2STAGE service. For the STILT computations all required input data as well as the output data are stored on one (or more) VM(s) in the EGI Federated Cloud and served to all VMs through the EGI DataHub providing access to the Open Data Platform, which supports management of distributed data. STILT model output is displayed on the web interface and users can download the results to their local computers. The web interface also accesses the data files stored on the Open Data Platform. Finally, all model output is

merged with output produced in previous model runs and the datasets are regularly transferred using B2STAGE to be archived in B2SAFE for long term storage.

Data requirements of the Footprint Tool are summarized in the tables below:

Input dataset		no of files per year	File size	Storage per year	Total storage 10 years
Emissions	EDGARv4.1	3	6.3 GB	19 GB	0.19TB
	EDGARv4.3	250	1-10 MB	< 1 GB	0.01TB
	VPRM	10	1-6MB	< 1 GB	0.01TB
	others	3-10	20-100 MB	< 1 GB	0.01TB
Boundary		5	1-10 GB	20 GB	0.20TB
Meteo		12	17G	205 GB	2TB
Particle location per station all available	1-3 hourly	2920-8760	0.5-5 MB	~20 GB	~ 20TB
	90 stations	~ 400000	0.5-5 MB	~2TB	

Output dataset		no of files per year	File size	Storage per year	Total storage 10 years
Footprints per station all available	1-3 hourly	2920-8760	40 kB	< 300 MB	~150 GB
	90 stations	~ 400000	40 kB	~15 GB	
user requests	new sites			~ 20GB	~200 GB
Time series per station all available	1-3 hourly	1	1-10 MB	~ 10 MB	~ 10GB
	90 stations	90	1-10 MB	< 1 GB	
user requests	new sites			10 GB	~100 GB
Particle location per station	1-3 hourly	2920-8760	0.5-5 MB	< 50 GB	------
user requests	new sites			500 GB	< 20TB

2.2.2 Implementation status

The implementation of EGI and EUDAT services and their interaction in the ICOS use case is largely based on the example of the Universal Use case. A X.509 robot certificate associated to ICOS CP was acquired for authentication and authorization at EGI and EUDAT services and for the testing

phase registered at the fedcloud.egi.eu Virtual Organization. The basic functionalities of the EGI and EUDAT services used in the different components of the footprint tool have been tested separately, and their integration into the workflow of the footprint tool is underway. While in the first set up the data required for the STILT computations is stored on block storage directly attached to the individual VM, in the next phase the data is stored on the central EGI Opendata Platform that is globally accessible, also from additional VMs if needed.

2.2.3 Next Steps

The full integration of the data management using EUDAT B2STAGE and B2SAFE services into the workflow at ICOS CP and in the footprint tool will be the next step. In addition, a strategy to attach persistent digital identifiers to the output data produced in the STILT computations and stored in B2SAFE will be developed.

2.3 EPOS

The *European Plate Observing System* (**EPOS)** aims at creating a pan-European infrastructure for solid Earth science to support a safe and sustainable society. In accordance with this scientific **vision,** the **mission** of EPOS is to integrate the diverse and advanced European Research Infrastructures for solid Earth Science relying on new e-science opportunities to monitor and unravel the dynamic and complex Earth System.

The EPOS architecture is composed of connected technical and organizational elements (Figure 4). In this conceptual organisation the project presents the Computational Earth Science (CES) component as the coordination framework for all the activities linking the distributed services (ICS-D) with thematic domain-driven services. In the context of the activity described by this deliverable, The ICS-D are represented by the services provided by the EGI and EUDAT infrastructure.

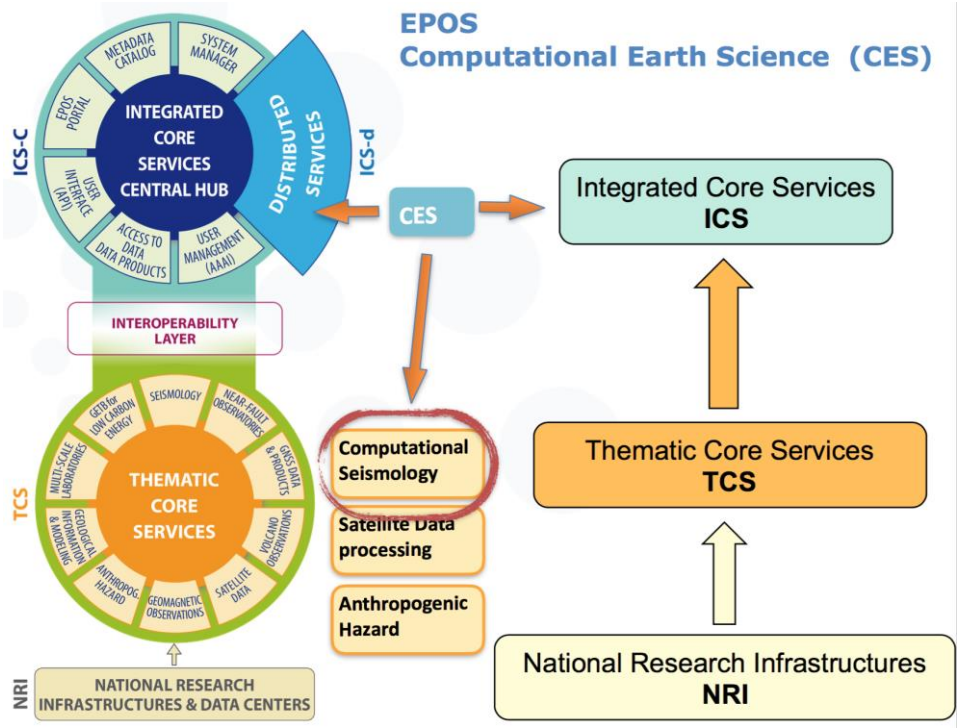

Figure 4 Key elements of the EPOS Functional Architecture and role of Computational Earth Science

The CES supports specific and science driven use cases according to the maturity of the specification and technical implementation. It aims and empowering the realisation of the use case with the adoption common standards and best practices offered by the e-science advancements.

Given these premises, we have selected a computational service in the field of Seismology as the target customer for the EGI-EUDAT integration. The choice has been conducted in coordination with the EPOS-CC use case after a careful evaluation of the competence centre activities.

2.3.1 Architecture

Figure 5 shows the main architecture of the target computational platform, which is better known as the VERCE (Virtual Earthquake and seismology Research Community e-science environment in Europe[3]). The system allows researchers to perform simulations real earthquakes, allowing moreover the misfit analysis of the synthetic data with real observations. It allows users to perform simulations adopting publicly available earth models, as well as experimental and therefore private ones. The Figure shows where the EUDAT and EGI services are expected to play a major role.

[3] http://portal.verce.eu

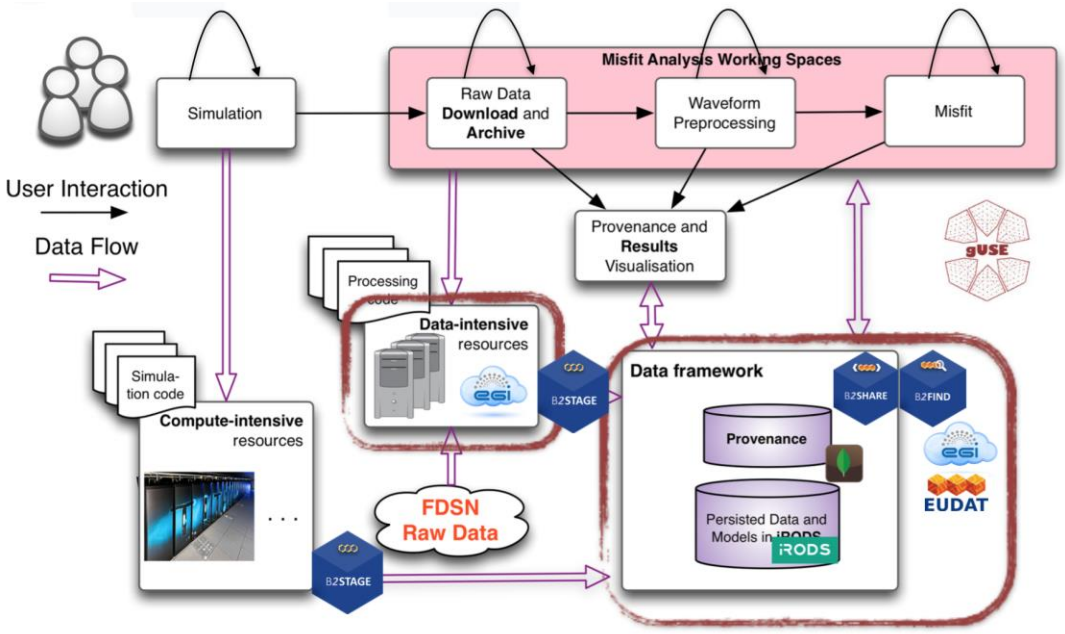

Figure 5 Computational Seismology Platform (VERCE): The image shows the interaction between the components of the VERCE platform and the impact of the EGI and EUDAT e-infrastructure in terms of computation (FedCloud), data-staging services (B2STAGE) dissemination

The simulation of a single earthquake performed on 500 cores for over 900 waveform channels measuring velocity, acceleration, displacement produces roughly 10 GB in 32 minutes. Misfit analysis includes a small part of this output (the seismographs). However, for a sensible computation of that would allow relevant results, the number of simulated earthquakes should be high, hence a full misfit iteration requires order of millions of waveform channels of 100KB each. The VERCE platform must be able to handle the transfer of Terabytes of data stored at EUDAT services coming from simulation, data access and processing workflows, which are executed at different stages. The metadata, provenance and medium-long term storage of all these files are crucial for the correct operation of the platform.

2.3.2 Implementation Status

This section describes the progress achieved and the motivations according to two major aspects of the computational service.

Enactment and Execution:

The VERCE solution, based the gUSE/WS-PGRADE system, is able to make use of different sort of resources, thus fostering sustainability and standardisation. We have started to implement all the required steps for including in the portfolio of resources the EGI Federated Cloud. As shown in Figure 5, we aim at using the EGI Federated Cloud to support data-intensive tasks, such as data pre-staging, processing and misfit analysis. We leave the simulation to be performed within classic HPC system SCAI, mostly because it guarantees a stable and reliable version of the parallel

SPECFEM3D solver software. Thus, to achieve the integration of the cloud resources we had to upgrade the gUSE/WS-PGRADE system to the latest version (v3.7.3), which is finally offering the compatibility to the OCCI interface.

Data Management:

The current iRODS[4] data-management system has been upgraded to v4 and migrated to a more stable environment at SCAI. This is a self-managed setup which does not belong yet to the EUDAT production infrastructure. We have started with the integration of the B2STAGE software stack on the server side as our data-movement mechanism allowing GridFTP secure transfers within the private user-managed data-space.

2.3.3 Next Steps

This is mostly given by the requirement of offering within the same data-management layer flexible web based functionalities for the visualisation and access to the data, besides the support of back-end services such as GridFTP transfer and the medium-term preservation of the experimental data produced by the platform. Basically, the current iRODS based data-store offers and combines what B2DROP, B2SHARE and to some extent also B2SAFE and B2STAGE would do as independent and distributed services. Improvements of the inter-communication between these three services are currently ongoing within the EUDAT2020 initiative. We hope in the next future to migrate the full data-management responsibilities to EUDAT facilities, once a complete integration of the B2 services is achieved.

The simulation results can be potentially large (ca. 20GB per simulation) but not always worth to be shared or officially identified by a PID. This brought improvements also to the GridFTP support for data-staging operations to and from iRODS, thanks to the integration of the EUDAT B2STAGE technology.

[4] http://irods.org

3 AAI Integration

The integration of the AAI of EGI and EUDAT would allow end-users to transparently access EGI and EUDAT services with the same credentials, so when a user is authenticated once on EGI and/or EUDAT, s(he) should be able to see the EGI and EUDAT services as offered by a unique infrastructure. Up to recently, EGI and EUDAT based their AAI on X.509 certificates and providing SSO was a matter of registering a certificate's Distinguished Name (DN) identifying the user on both infrastructures to achieve the integration. However, both infrastructures are progressively migrating their AAI to federated authentication mechanisms based on SAML and/or OpenID Connect that remove the need of users to handle X.509 certificates and enable the use of home institution identity of the user in the infrastructure. In the new schema, both EGI and EUDAT services for AAI follow a similar architecture with a central element (EGI CheckIn for EGI and B2ACCESS for EUDAT) that acts as identity proxy between the user and the resource providers and handling the complexity of multiple IdPs/Federations/Attribute Authorities/technologies.

The integration of the new AAI systems of EGI and EUDAT started on May 2016 with an initial meeting to describe the AAI architecture of each infrastructure and the definition of an activity plan to make them interoperable. The plan identified the following steps to progressively achieve the integration:

- Allowing users to access EGI and EUDAT web services with the same credential. By registering EGI CheckIn in EUDAT B2ACCESS and vice versa a user from one infrastructure can login with his/her identity on the other system.
- Allowing users to access EGI and EUDAT non-web services with the same credential. In this case, once the user is logged in, token translation services will provide a X.509 certificate specific for the infrastructure that needs to be accessed.
- Attributes harmonisation. EGI CheckIn and B2ACCESS provide a set of attributes that allow the resource providers to take authorisation decisions when the user wants to access them. These should be harmonised to avoid providers having to configure specific rules for users of each infrastructure.
- Enabling EGI services to delegate user's credential to EUDAT services and vice versa. The delegation would enable true SSO since the user would not need to login to both CheckIn and B2ACCESS but would transparently be able to access services with his/her initial credentials.
- Data privacy issues and policy harmonisation, for sharing attributes between the EGI and EUDAT infrastructures and following AARC guidelines on policy harmonisation.

A follow-up meeting was held on January 2017 to start with the integration process as defined above. Currently tests are ongoing for the registration of CheckIn into B2ACCESS and vice versa.

4 Plan for Exploitation and Dissemination

4.1 ICOS

Name of the result	ICOS Carbon Portal: Footprint Tool
DEFINITION	
Category of result	Software & service innovation
Description of the result	Online tool to analyse potential contributions of natural fluxes and anthropogenic emissions to the atmospheric CO2 concentrations at a selection of ICOS atmospheric stations
EXPLOITATION	
Target group(s)	Researchers, especially people (within ICOS or planning to join) that own an observation station.
Needs	Understanding the "footprint" of the ICOS observation stations.
How the target groups will use the result?	In further research activities other than those covered by the project concerned
Benefits	Interactive, on-demand service that computes and visualizes the sensitivity of GHG concentration signals at potential and existing ICOS atmospheric measurement stations to GHG emissions and fluxes from different sources
How will you protect the results?	Open Source License: GPL 3
Actions for exploitation	Deployment of the Fooprint tool into production. Integration into the ICOS Carbon Portal Development of documentation and training material
URL to project result	https://stilt.icos-cp.eu/
Success criteria	Number of users of the footprint tool Number of simulations executed
DISSEMINATION	
Key messages	Interactive and on-demand calculations of measurement station footprints. Combine data on greenhouse gas emissions and uptake with meteorological

	transport models.
Channels	Presentations on conferences and scientific publications
Actions for dissemination	Poster at the DI4R Krakow 2016 Presentations at EGI conferences in 2015 (Bari) and 2016 (Amsterdam) Scientific background presented at various reports and conferences
Cost	
Evaluation	Number of users of the footprint tool Number of simulations executed

4.2 EPOS

Name of the result	VERCE
DEFINITION	
Category of result	Software & service innovation
Description of the result	Data-intensive e-science environment to enable innovative data analysis and data modelling methods that fully exploit the increasing wealth of open data generated by the observational and monitoring systems of the global seismology community
EXPLOITATION	
Target group(s)	Seismology researchers
Needs	Understanding Earth's internal wave sources and structures, and augment applications to societal concerns about natural hazards, energy resources, environmental change, and national security.
How the target groups will use the result?	In further research activities other than those covered by the project concerned
Benefits	Delivers a comprehensive integrated and operational virtual research environment (VRE) and e-Infrastructure to harvest the new opportunities provided by data-driven high-performance full waveform simulations (FWS), together with a data-intensive framework for collaborative development of innovative statistical data analysis methods. All accessible from a single endpoint, the VERCE science gateway.

How will you protect the results?	Open Source License: MIT
Actions for exploitation	Deployment into production of the developments on the VERCE portal Development of documentation and training material
URL to project result	http://portal.verce.eu
Success criteria	Number of users of the portal
DISSEMINATION	
Key messages	VERCE is a unique seismological web platform will allow seismologists to archive, access and analyse a massive amount of data thanks to innovative applications
Channels	Presentations on conferences and scientific publications Social networks (LinkedIn, YouTube)
Actions for dissemination	Presented project in partners websites Prepared two posters, one general and one presented at ESC Moscow 2012 Published YouTube videos and LinkedIn group posts Published of posts from Facebook and Twitter 15 Scientific publications and one book chapter released Detailed list of activities is available at VERCE web : http://www.verce.eu/PublicDissemination.php
Cost	
Evaluation	Number of users of the VERCE portal

5 Conclusions

The EGI-EUDAT integration activities have demonstrated that use cases can leverage services from both infrastructures in a single application. The universal use case has proven that basic interoperability is possible and served as a basis to build more complex use cases such as the ones selected after the cross-infrastructure case studies call: ICOS and EPOS from the earth observation community. These are progressing well towards a full production deployment that can serve researchers in their respective areas. Monthly meetings to keep track and of the developments and identify any potential blockers are regularly held with representative from both use cases. Future plans for these use cases will focus on the progressive inclusion of services from EGI and EUDAT. The experience gained with the ICOS and EPOS use cases will serve as guidance for new use cases and to the expansion of the universal use case covering more services.

AAI Integration is a key cornerstone for the interoperability of EGI and EUDAT e-infrastructures. While integration with X.509 is already possible as demonstrated by the selected use cases, the use of federated authentication mechanisms will be gradually introduced in the e-infrastructure and a plan for the interoperability at that level has been also defined. First steps for the integration (registration of EGI CheckIn into EUDAT B2ACCESS and vice versa) are ongoing with tests from the AAI teams of each e-infrastructure. Once this activity is finalised, users from either EGI or EUDAT will be able to transparently access the resource providers of any of the infrastructures with a single credential.