14 June 2017

Dear members of the EGI-Engage Collaboration,

With this issue of the EGI-Engage letter I would like to update you on the progress of our work at a time when the project is reaching its final stage.

First of all, Amendment 2 is now signed by the EC and the documentation is now available for your consultation.

Please note that additional changes of budget distribution to compensate underspending and overspending will be implemented at the end of the project before the final payment. No additional amendment is envisaged.

Meeting many of you at the EGI Conference and INDIGO Summit 2017 was a pleasure. The conference offered the opportunity to share information about the progress of activities, accelerate the implementation of plans, and develop existing and new collaborations with partners and sister projects.

Our next appointment will be in Brussels for the Digital Infrastructures for Research conference DI4R 2017. The conference will take place on 30 Nov 30 – 01 Dec, and will be co-located with the EOSCpilot stakeholder conference (28 and 29 Nov).

This letter continues with a report of the main outcomes of our project from March 2016 till February 2017. I hope you find it useful to get a bird-eye view of our accomplishments!

Yours sincerely,

Tiziana Ferrari, EGI-Engage Technical Coordinator

EGI-Engage outcomes

In February 2017 EGI-Engage reached the end of its second year. A number of milestones have been achieved by EGI and partners.

Services for open science.

Today EGI is an initiative of 23 partners: 22 NGIs and one international research organization – CERN. With the support of EGI-Engage and the Council, we completed the definition of the EGI internal and external services catalogues. This was an important milestone to develop a clear value proposition, increase the accessibility of available services, and make the providers from the EGI federation more visible to funders and policy makers. The external catalogue includes services for researchers and is organized across three broad categories: Compute (including Cloud Compute, Cloud Container Compute and High-Throughput Compute), Storage and Data (Online Storage, Archive Storage and Data Transfer) and Training (FitSM and the Training Infrastructure).

On the other hand, the internal catalogue offers the services necessary to federate local, regional and national infrastructures into the **EGI Federation**. They provide the "glue" that transforms individual services into a harmonized pan-European system for international research collaborations, and include to name a few: cross-domain accounting and monitoring, identity provisioning, federated authentication and authorization, security coordination, and user support.

On top of the internal and external catalogues, the [EGI Scientific Applications](#) provide ready-to-go access to the EGI Federation. They offer community-specific capabilities like datasets, search capabilities, data analytics, and scientific software, and are provided by many partners like CREATIS, CNRS, the National Bioinformatics Infrastructure of Sweden, Terradue, and the WeNMR research community to name a few. Scientific applications are being used by thousands users worldwide and are a tremendous research enabling instrument. We are looking at expanding the scientific application portfolio with contributions from the NGIs leveraging their national efforts, the EGI-Engage Competence Centres and any other partner who is willing to engage with EGI.

To make these services discoverable and accessible online, the project has been successfully prototyping a Marketplace, which we are planning to bring in production in August. The marketplace will give the possibility to browse services by category, choose configuration options and place service orders.

Infrastructure

The High-throughput Compute infrastructure comprehends to date 635,00 logical CPU cores and in the last 12 months the compute power provided increased by +15%. The estimated storage infrastructure amounts to 540 PB (+13%) and the compute workload collectively supported by the infrastructure in the last 12 months increased by +10%.

The EGI Federated Cloud has been developing into a multi-national cloud system with the contribution of 24 providers. It integrates community, private and/or public clouds into a scalable computing platform for research. The Federation pools IaaS, PaaS and SaaS services from a heterogeneous set of cloud providers using a single authentication and authorization framework that allows the portability of workloads across multiple providers and enable bringing computing to data. Thanks to the federation of cloud compute capabilities, user communities can easily analyse data in their own cloud infrastructure. By the end of EGI-Engage, we will work to improve its reliability, we will port different platforms that enable the federated distributed access and we will provide new mechanisms for cloud management,

(such as the AppDB VO Operations Dashboard) In addition, OpenID Connect integration will be completed and the EGI DockerHub of trusted containers will be offered.

Cloud-based user communities.

The EGI scientific applications ported to the cloud collectively serve to date more than 15,000 registered users. In the context of a recent collaboration, EGI-Engage has been supporting the EMSODEV project of the European Multidisciplinary Seafloor and water column Observatory (EMSO). The EGI Federated Cloud providers - RECAS-BARI, INFN-Padova, CESGA, LIP – enable the running of the EMSODEV Data Management Platform (DMP), which addresses the needs of biologists, geoscientists, chemists, and engineers. DMP has been designed to exploit capabilities of e-Infrastructures to develop a flexible and scalable Data Management Service for a long-term, high-resolution, monitoring by providing a coordinated approach for data capture, archiving, management and delivery based on OGC standards.

Another example of cloud integration is offered by the ELIXIR use cases. Thanks to the ELIXIR Competence Centre, in the *ELIXIR Compute Platform rOCCI, jOCCI, OpenStack Cli and Terraform can now be used to manage ELIXIR Virtual Machines (VMs). CESNET, EMBL-EBI and GRNET are providing resources to the ELIXIR Virtual Organization.* Three out of five scientific use cases have effectively used EGI Federated Cloud: Marine Metagenomics, Insyght comparative genomics and PhenoMeNal.

With the support of the dedicated competence centre, **LifeWatch has been the first Research Infrastructure embracing federated cloud compute as their production-ready compute paradigm**. The LifeWatch Competence Centre facilitated the exploitation of available solutions by the LifeWatch user community by providing basic tools for data management, data processing and modelling, tools for Ecological Observatories, Virtual Labs. LifeWatch has been enabling the direct participation of citizens by allowing them to contribute observation records and offering compute services. The LifeWatch applications running partially or totally on federated cloud compute are: the Collaborative Platform for observatories, Water Mass modelling, the GBIF data access biogeographic context, Image Classification Deep Learning Tools, Citizen Science tools, R services, and the Digital Knowledge Preservation Framework.

The integration of D4science with the EGI Cloud Federation was completed. Thanks to this, D4science will use cloud resources to scale up internal compute facilities to support various research disciplines. To date, D4science provides 91 VREs to marine science, social data mining, environmental research, cultural heritage, agriculture and digital libraries. This integration will allow a range of different VREs to be run of the EGI Federated Cloud.

[The EGI Federated Cloud blueprint](#) **is being promoted widely in the context of the "Boston Open Research Cloud Declaration"**. The initiative aims at exchanging policies and practices that facilitate interoperability and shared use of data and compute resources between and among national and institutional research clouds. The initiative will also develop a community declaration of those guiding principles.

New services

Thanks to the Open Data Platform testing activities, the **DataHub** is currently being validated to address two use cases: the cross-domain federation of existing storage infrastructures, and the provisioning of a distributed platform for the management of replicas of publicly data collections. Once in production, the DataHub will increase accessibility to third-party research data for downstream analysis via federated authentication and authorization, the bridging of preserved data and computing with the federation of

distributed data repositories and the possibility of associating permanent identifiers to the output data that is generated by the processing workflows of EGI.

In order to support collaborative science, we have launched the **Applications on Demand (AoD)** service to offer a fully-featured, scalable High Throughput Computing environment for research and scientific computing. AoD delivers a whole cluster complete with job scheduler and a library of scientific applications and other utilities, with the possibility to store and retrieve the output data, workflows and results. The applications and tools offered to date include: thematic applications for supporting Life Sciences disciplines (Galaxy, ClustalW2, Chipster, NAMD and AutoDock Vina); generic utilities (Docker, Apache Tomcat, Hadoop, Marathon, and Chronos); applications for Engineering (GnuPlot, Octave and the Statistical R for Computing and Jupyter Notebook) and Art and Humanities (the parallel Semantic Search Engine). In addition to the library, AoD allows researchers to run custom scientific applications. Underpinning compute and storage services are currently provided by CYFRONET, INFN, BELSPO, MTA SZTAKI, CESGA, Universitat Politecnica de Valencia, and BIFI. All NGIs interested in making the AoD library available to national user communities can simply plug their national IaaS resources to the service. Through AoD: (1) researchers can access a ready to use library of applications underpinned by a compute and storage infrastructure, to carry out compute and data-intensive analysis, (2) developers can integrate custom applications into the service and share them with other researchers, and (3) NGIs and other providers of compute and storage services can plug their application library and/or local clusters and clouds to serve their national user communities without having to manage a in-house system for the identification, authentication and authorization of users.

A ground-breaking achievement of PY2 of EGI-Engage is **CheckIn.** CheckIn is a proxy service that operates as a central hub, connecting federated Identity Providers (IdPs) residing outside the EGI ecosystem, and Service Providers (SPs) that are part of EGI.

Via CheckIn users are able to authenticate with the credentials provided by the IdP of their home organisation (e.g. via eduGAIN), as well as by using social identity providers, or other selected external identity providers (support for eGOV IDs is also foreseen). To achieve this, the EGI AAI has built-in support for SAML, OpenID Connect and OAuth2 providers and already enables user logins through Facebook, Google, LinkedIn, and ORCID. In addition to serving as an authentication proxy, the EGI AAI provides a central discovery for users to select their preferred IdP. CheckIn is also responsible for aggregating user attributes originating from various authoritative sources (IdPs and attribute provider services) and delivering them to the connected EGI service providers in a harmonised and transparent way. Service Providers can use the received attributes for authorisation purposes, i.e. determining the resources the user has access to.

CheckIn was successfully integrated with the ELIXIR AAI infrastructure, which operates both an IdP and attribute provider service to manage user accounts and personal attributes for every ELIXIR user. During the last months the EGI AAI proxy has been integrated with the ELXIR AAI with the goal to enable members of the ELIXIR Community to access EGI services. Two EGI services – the GOCDB service registry and the AppDB Virtual Machine marketplace – were chosen as priority services for the ELIXIR integration. The integration work was achieved by connecting the SP proxy element of the EGI AAI proxy with the IdP proxy of the ELIXIR AAI by exchanging their SAML metadata in XML format. Besides these, OpenStack was also integrated. As a result of this work ELIXIR can interact with GOCDB, AppDB and OpenStack resources using their ELIXIR user identities. In addition to this, CILogon can be used by ELIXIR users to create proxy certificates.

Among the new compute platforms being brought into production, **GPGPU clusters** are getting momentum. WeNMR – the world-wide virtual research community supporting structural biology – after having successfully benchmarked various tools of use in structural biology applications, implemented GPGPU-enabled portals for some of them. Specifically, they implemented GPGPU-enabled web interfaces for the AMBER package for molecular dynamics simulations and for the DisVis tool for quantifying and visualizing the accessible interaction space of distance-restrained binary complexes. These new portals provide an enhanced service to the community, by exploiting the significantly faster performance achieved on a GPGPU computational infrastructure. Since August 2016, DisVis and Powerfit were adopted by 70 and 59 users respectively. For the AMBER package, the portal can automatically define the most appropriate computational infrastructure to dispatch the calculations, based on the protocol selected by the user. DisVis can however also run in multicore mode instead of GPGPU. In the future WeNMR will also investigate use of GPGPU-enabled cloud resources.

The **Accounting Portal** was reengineered and now it presents new features and a renewed look and feel based on the Django Python framework, the Dojo and Bootstrap Javascript frameworks. In addition, the first prototype for gathering data usage accounting records was developed. This new feature enriches EGI accounting by allowing data providers to get information about the reuse of datasets hosted in the EGI infrastructure.

The implementation of the **EGI Service Integrated Management System** (IMS) **was completed and ISO 9001 and ISO 20000 certification were obtained**. The EGI IMS – compliant to the FitSM standard – defines the activities performed by the providers of the EGI Federation to plan, deliver, operate and control the EGI services offered to customers. The activities carried out are directed by policies and are structured and organised by processes and supporting procedures. The IMS provides a new structure to the existing operational activities, procedures and policies, and ensures that all components are regularly reviewed, properly documented and roles and responsibilities are clearly defined. The IMS is organized in three parts: the **Core Management System** (to plan, implement, monitor and continually improve all business processes), **General Service Management** (to deliver all services covered by the EGI service catalogue) and the **IT Service Management** (to deliver IT services).

EGI-EUDAT service integration. The first project outcome has been the definition of a universal use case that covers generic user needs. Integration activities have demonstrated that use cases can leverage services from both infrastructures in a single application. The universal use case proved that basic interoperability is possible and served as a basis to build more complex use cases such as the ones provided by ICOS and EPOS. These are progressing well towards a full production deployment that can serve researchers in their respective areas. AAI Integration is a 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. The registration of EGI CheckIn into EUDAT B2ACCESS and vice versa is ongoing.

Competence Centres. Besides the achievements of ELIXIR and WeNMR, the competence centres progressed with their service integration and proof of concept activities.

BBMRI. Human biobanks are the core of BBMRI-ERIC medical research infrastructure; these are repositories of biological material and data associated with the research participants (donors or patients willing to participate in the research). The associated data covers a broad range of data types: from data collected directly from the research participants and medical processes related to them, to data generated from the biological material. BBMRI CC worked on describing biobank data processing workflows that were selected for piloting in EGI-Engage with BBMRI.nl and BBMRI.cz together with their associated biobanks. The main focus of these is on proteomics and genomics workflows, which cover both extremes of privacy-sensitive data processing spectrum: from relatively non-sensitive applications to very sensitive ones. In addition, the AAI components of BiobankCloud – a BBMRI platform which builds on a Hadoop distribution called Hops to provide scalable storage and processing for genomic and Biobank data –was extended to support federated authentication with Shibboleth.

Disaster mitigation. Two simulation portals were developed by the Disaster Mitigation Competence Centre (DMCC) during EGI-Engage: tsunami wave propagation simulation portal (iCOMCOT) and weather simulation portal (gWRF). The two portals provide compact and ease-to-use simulation tools for the full cycle of a tsunami event and numerical weather prediction. The portals are now open for researchers and will be used by the Disaster Mitigation Competence Centre of the EGI-Engage project to simulate storm surge, flood, and tsunami and forest fire dust transportation events from Philippine, Malaysia, Thailand and Taiwan. In parallel with this, Academia Sinica is continuing the development of the portals.

DARIAH. A science gateway was developed to provide convenient web-based access to the resources available in the DARIAH CC VO. The gateway hides the details of accessing the compute and storage resources, and offers user-friendly portlets for running tasks related to the arts and humanities community. The portal currently offers three applications (Simple Semantic Search Engine, Parallel Semantic Search Engine, DBO@Cloud) and three services (Cloud Access, Workflow Development, File Transfer). The gateway provided by the DARIAH CC is based on WS-PGRADE/gUSE. The federated login enables researchers of eduGAIN member institutes to access the gateway by using their home institute credentials, thus there is no need to maintain a separate login

EISCAT-3D. The competence centre designed and developed a prototype of the computing and data model that will support the data management and analysis from the EISCAT ionospheric and atmospheric radar in the auroral zone in the Fenno-Scandinavian Arctic. EISCAT-3D adopted DIRAC as solution to get: a specific EISCAT_3D web frontend based in the DIRAC web framework, and the connection to distributed computing (HTC and cloud), storage and AAI. The prototype is now in pre-production stage. New features were added to the EISCAT_3D portal like data access based on the EISCAT data access policy, the data downloading was refined, as well as file submission.

EPOS. The competence centre helped extending the level of interoperability with some of the services of the EUDAT portfolio and conducted three pilot demonstrators highlighting how EPOS itself could benefit from EGI services for the delivery of the EPOS Distributed Integrated Core Services (ICS-D). In the context of the first pilot, the EPOS AAI infrastructure based on UNITY IDM was integrated with the EGI CheckIn service. The second pilot for earthquake simulation (MISFIT) showed how the back-end services of the application could be extended through the integration with the EGI Federated Cloud. Furthermore, EPOS defined various EGI-EUDAT service integration scenarios: software previously developed by the VERCE project was enabled on the EGI Federated Cloud, using data from EIDA/ORFEUS organization via B2SAFE

and B2STAGE. Finally, the third case – linked to the EPOS satellite data Thematic Community Service – demonstrated the adoption of cloud compute services for the analysis of Earth Observation datasets collected by various satellites, including the Sentinels of the Copernicus programme. The Geohazard Thematic Exploitation Platform, developed by Terradue on behalf of ESA, today exploits EGI Compute and Storage services for P-SBAS InSAR Sentinel-1 TOPS developed by CNR-IREA. P-SBAS stands for Parallel Small BAseline Subset and it is a DInSAR processing chain for the generation of Earth deformation time series and mean velocity maps, using SLC (Level-1) Sentinel-1 input data.

During PY2 security operations advanced significantly. A new security threat risk assessment was performed, focusing in particular on the security of the EGI Federated Cloud service. A security risk mitigation in the EGI FedCloud was to require the use of endorsed/approved VM images or Virtual Appliances. Another area of concern was the difficulty of detection and handling of security incidents in the EGI FedCloud as were threats arising from vulnerabilities and their handling.

Many developments have been made to the EGI CSIRT/IRTF procedures. These included the *EGI Security Incident Handling Procedure* and forensic guidelines. We have improved the *EGI-CSIRT Critical Vulnerability Handling procedure*, to be more readable and accessible and to address issues related to the new concept of Virtual Appliances. IRTF developed internal tools in order to decrease the number of repetitive tasks and to standardise the different messages sent.

Plans for an enhanced security challenge framework for the EGI FedCloud were agreed by the EGI CSIRT. The framework has been developed and tested at a number of sites.

New security policies have been produced and others have been revised and updated. The changes made were to address the evolution of EGI services and technology and to mitigate risks identified in the new security risk analysis. This included a major revision to the top-level security policy document; an updated version of the "Acceptable Use Policy", generalised to include all EGI service offerings. A complete revision of the policy on Data Protection was performed to turn this into a more general framework to be used by all services. Other policy work included a new policy on Acceptable Authentication Assurance and updating of the terminology used by all remaining old security policies that needed no other updates.

The EGI Software Vulnerability Group carried out a major revision of the Software Vulnerability Issue handling procedure to address the evolving EGI services. For example, when a critical vulnerability is found relating to a Virtual Appliance, this needs to be updated urgently by those responsible for the VA.

In the **WISE community** through its representation in the steering committee and leadership of working group activities, EGI benefits from better alignment of security practices and increased input into its risk assessment, training and trust programmes. EGI leads the SCIV2 working group, where work is ongoing on a new version of the SCI trust framework. The construction of comparative policy frameworks allows easy movement of researchers and data across multiple Infrastructures, whilst the construction of the human network through periodic WISE workshops has also established communication channels for operational security activities.