

Inspired

ISSUE 21
OCTOBER 2015

news from the EGI community



TOP STORIES

New standard for IT Service Management

page 2

From FedCloud to Open Science Cloud

page 3

EGI and AARC - the future of AAI

page 6

A new name for EGI

page 1

MORE

- 06 Data Science Professional, uncovered
- 07 PhenoMeNal - a new Life Science e-Infrastructure
- 08 New project: West-Life
- 09 Community Forum 2015: Data without boundaries
- 10 Community Forum 2015: Preview and highlights



Engage - Grow - Innovate

www.egi.eu

This Issue

In this edition of *Inspired*:

- > **Sy Holsinger** writes about the final outputs of the FedSM project and a new standard in IT Service Management
- > **Tiziana Ferrari** writes about the Open Science Cloud, a new idea high up in the European agenda
- > **Peter Solagna** gives an update on the progress of the AARC-EGI collaboration
- > **Andrea Manieri** and **Małgorzata Krakowian** tell us how the EDISON project will contribute to define a career for Data Scientists
- > **Namrata Kale** and **Christoph Steinbeck** introduce PhenoMeNal, an e-Infrastructure for pheno- and genotyping data
- > **Chris Morris** writes about West-Life, a new project set up to support structural biology
- > **Sergio Andreozzi** introduces Data without boundaries, the data track at the Community Forum
- > And we give an overview of the upcoming Community Forum in Bari



Our next event will be held in the port city of Bari in the Italian coast of the Adriatic Sea. (Image: Tango7174 / Wikimedia Commons)

Your feedback and suggestions are always welcome!

A new name for EGI

EGI updated its name and its statutes to keep up with evolving times

The EGI Council has decided to simplify the name of the European Grid Infrastructure to EGI – just the acronym without an expansion.

The change reflects the evolution of EGI towards becoming a worldwide, technology-agnostic e-Infrastructure for science.

The name change of European Grid Infrastructure to EGI is part of an update in the statutes of the EGI Foundation signed over the summer.

New statutes

The new statutes reflect changes in the governance of the EGI Foundation:

- > Full membership is now open to non-European countries – participation is no longer restricted to a geographical area

- > EIROs, ERICs and other organisations that support the objectives of EGI can also become full members from now on.

“The changes are meant to emphasize the international interest and mission of EGI, which are not limited by a particular technology or geographical area,” says Yannick Legré, Managing Director of the EGI Foundation. “We aim to serve excellent science, regardless of location and discipline.”

Matthew Dovey, Chair of the EGI Council adds, “updating the statutes of the foundation and the EGI name will reinforce cooperation with other pan-European and worldwide e-

infrastructures and research infrastructures, bringing us closer to the way science is actually undertaken.”

From now on...

EGI is just EGI

Full membership open to non-European countries

Full membership open to all research organisations

New standard for IT Service Management - FitSM

Sy Holsinger writes about the final outputs of the FedSM project

Over the last three years, EGI.eu and the National Grid Initiatives of Finland (CSC) and Poland (PLGrid) had the opportunity to participate in FedSM, a recently concluded EC-funded project. The project aimed at helping e-Infrastructures adopt service management best practices. The major result of the project was the development of a new lightweight service management standard called FitSM.

FitSM is fully compatible with other major ITSM standards and good practice frameworks such as ITIL and ISO20k. Its advantage is a much more pragmatic and achievable standard with minimum requirements around the delivery of IT services, as well as practical ITSM implementation guidance through templates, samples, guides and a formal training and certification scheme.

The training and certification scheme is backed by certification authority TÜV SÜD, and offers a range of course levels from Foundation, two Advanced (Service Planning and Delivery; Service Operation and Control) and an Expert level.

As a client partner in the project, EGI.eu, focused on the implementation of FitSM, first centrally, then moved across the infrastructure. The revision of the EGI service portfolio benefitted greatly from FitSM. The implementation led to a clear separation between what EGI.eu offers to the participant organisations, and what EGI, collectively as a federation, offers to individual researchers and research communities.

As a result of this exercise, we created a solutions' portfolio to better communicate the value delivered by EGI's services.

FitSM also provided the guidelines and templates to re-structure the entire framework of Service and Operational Level Agreements (SLAs and OLAs) in federated environments.

The net effect was increased clarity. Federations can be quite complex, so having good descriptions of services makes the value of EGI and what it delivers much clearer. Overall, the process orientation of FitSM helps drive a more reproducible and reliable activity on a daily basis, and provides a framework for connecting the different processes such as service level management, incident and service request, customer relationship, release and deployment, amongst others (14 FitSM processes in total).

The training and certification scheme helps to support this goal. So far, all technical staff at EGI.eu has received the Foundation training, half with the Advanced, another handful with Expert level and two experienced trainers. In addition, more than 150 members of the EGI community have received training as well.

A standard for the future

Moving beyond the project, FitSM is now maintained by a non-profit organisation called ITEMO (IT Education and Management Organisation) to which EGI and several project partners are members and



actively participate in the FitSM Working Group to ensure the long-term sustainability of the standard.

EGI.eu is also a registered training organisation with TÜV SÜD and has added FitSM Training and Certification to its service portfolio.

At the time the FedSM project started, EGI was just looking for support to help improve day-to-day activities. We never anticipated being a part of what FitSM is today and we feel quite proud to have contributed to the development of FitSM as the popularity of the lightweight approach is spreading throughout not just e-Infrastructures, but the larger ITSM community as well.

More information

FitSM:
<http://www.fitsm.eu>

ITEMO:
<http://www.itemo.org>

EGI FitSM Training Service:
<http://go.egi.eu/fitsm>

From Federated Cloud to Open Science Cloud – reflections on the future of clouds and data

Tiziana Ferrari writes about a new idea high up in the European agenda

In May 2015 the European Competitiveness Council welcomed the development of a **European Open Science Cloud** to enable sharing and reuse of research data across disciplines and borders, taking into account relevant legal, security and privacy aspects.

This offers an opportunity to reflect both on the experience gathered through the development of the EGI Federated Cloud and on the new challenges and requirements from the seven Research Infrastructures cooperating with EGI in the context of EGI-Engage. What has been achieved so far and what is missing to make the Open Science Cloud come true?

After 18 months of operations, the EGI Federated Cloud brings together 22 cloud providers across 15 European countries. So far, the infrastructure provided over 800,000 Virtual Machines to research communities.

What has been proven to be a success and what remains to be addressed?

A few thoughts.

First, the **Applications Database** for virtual appliances has proven to be a powerful tool. The AppDB Cloud Marketplace allows the instantiation of VMs on a network of cloud providers that adopt open cloud standards and support federated AAI. This trust model assures cloud providers that virtual appliances in the library are endorsed by research groups (the so-called Virtual



Organisations). The availability of a community service for the sharing of tools, community platforms, scientific codes and research data, is being increasingly advocated to be one of the enablers of Open Science. In a recent meeting on data-driven research, held at the RDA plenary in Paris, Jan Korbel – member of the steering committee of the Pan-Cancer project – advocated the availability of a virtual marketplace for the aggregation and analysis of genomes and associated data would secure the future of cancer genomics. He argued that in exchange for depositing data, users could get prioritized access to point-and-click menu of bioinformatics tools to run on data.

This vision advances the AppDB by establishing a framework for open access, managed access policies and IPR management. The **federated approach to IaaS** is also key. By providing a distributed hosting environment,

community-specific web services can be instantiated in different cloud providers. Virtual appliances can migrate across the federation to provide resiliency and portability and, most importantly, cloud compute services can be offered where big data resides, without the need of moving research data out of the institute premises, a key requirement in case of sensitive data.

Portability of applications and data requires the adoption of **open standards**. As cloud providers are heterogeneous across a federation, EGI is supporting different interfaces of choice but promotes the adoption of OCCI and CDMI for full interoperability. However, standardization is necessary at all levels of the platform infrastructure. Community standards are equally necessary to allow discoverability, use and reuse of data, as reported at the RDA meeting by David Schade from the Canadian Astronomy

What are the new challenges the Cloud faces to support Open Science?

Federation, open standards adoption, scalable access to research data, portability and sharing of tools and software

Data Centre, an organisation that provides data services to more than 15,000 researchers worldwide thanks to its data sharing and open data policies. The question that arises is: what do we need to scale up capacity and capabilities to enable international collaborations?

How should EGI evolve?

1. A federated approach to research data

Nowadays, research practice is increasingly and, in many cases, exclusively data-driven. Knowledge of how to use tools to manipulate research data, and the availability of e-Infrastructures to support them, are fundamental. New types of communities are forming around interests in digital tools, computing facilities and data repositories. By making infrastructure services, community engagement and training inseparable, existing communities can be empowered by new ways of doing research, and new communities can be created around tools and data.

The Open Science Cloud should provide a marketplace making open research data, the related tools and knowledge discoverable. The marketplace would federate existing research data sets that are provided by archiving organizations that can ensure compliance to a set of quality standards defined by the marketplace, and provide tools and computation services. As Jesús Marco de Lucas, from the Spanish IFCA-CSIC and LifeWatch, explained to the RDA audience, this approach would

not just benefit Research Infrastructures but also citizen scientists.

2. Offering of scalable access to and analysis of research data for use and reuse

Making data findable is not enough. Researchers should not need to download large datasets locally before executing their workflows – this would make access to data less efficient and time consuming. Research data must be made easy to access and reuse and this means scalable access, especially in case of big data that cannot be efficiently downloaded locally. The Open Science Cloud could provide distributed data mirroring and caching capabilities based on federated IaaS cloud storage.

This would allow the Open Science Cloud to realize the Data Commons, extending practices already adopted by a number of research communities (e.g. in genomics) to all disciplines. This is a key pillar of the Open Science Commons vision that EGI is promoting.

3. Integrating (shared) tools and applications

Knowledge cannot be extracted from data without specialized tools and applications (e.g. for text mining). The Open Science Cloud could provide a library of

community-specific applications and tools that should be open for publishing to any researcher. For greater specialization, the Open Science Cloud should provide PaaS and SaaS services that are community-specific and that could be dynamically deployed with a focus on the long tail of science. These services could be provided in the form of managed services by the Research Infrastructures.

4. Provide services for depositing data for resource-bound users

The Open Science Cloud should federate infrastructures to provide services for the long tail of science, in particular for those researchers that cannot benefit from these services at institutional and/or national level, so that the digital divide that still concerns various European regions can be mitigated.

5. Achieving integrated e-infrastructures

The development of an Open Science Cloud with an inclusive federated governance model avoids duplication of provisioning of ICT services at national and European level. The Open Science Cloud should be developed as a federation of national cloud hubs, financially supported by member states. The role of the EC would be to



ensure the persistency of the services that allow the national cloud hubs to operate as a federation, and to ensure the coordinated procurement, service provisioning and data brokering according to the requirements of the RIs. This would allow aggregation of demand across Europe, coordinated delivery and the development of economies of scale.

Why a cloud?

Cloud is a service provisioning paradigm that can host data and software tools.

Being based on virtualization, clouds facilitate sharing, reuse and the combined offer of data and tools. Cloud federations enable 'local hosting' and 'control sharing' capabilities to respect ownership and allow accessibility for distributed communities. In addition, the federation approach allows the implementation of hybrid models where private, community and public clouds can be integrated.

How do we organise an Open Science Cloud?

The Open Science Cloud could be managed as a federation of cloud hubs, jointly provided through a hybrid approach, involving publicly-funded and commercial cloud providers. A

federation of hubs provides an organizational structure that meets European policies, regulations, restrictions and business models, which in some cases do not make the permanent relocation of data into centralized science-domain specific repositories possible, and/or into generic repositories (that integrate data and tools from multiple domains). Within this federation data providers are able to retain complete control to their data.

The federated approach allows the implementation of a multi-level governance model where different governing bodies of the Commons can coexist and integrate.

Who should operate the service offered by the cloud hubs?

Cloud hub services should be coordinated across multiple stakeholders, including research communities, research infrastructures and e-Infrastructures. A federator role needs to be established to ensure services are provided, via a single interface and according to federated service management best practices and standards.

EGI is keen on engaging with Research Infrastructure, research communities and other e-Infrastructures to jointly discuss these issues.

Open Science Cloud workshop

If you are interested in bringing your view on the Open Science Cloud do not miss the Community Workshop on the Open Science Cloud "Shaping the Open Science Cloud of the Future" promoted by EGI, GEANT, EUDAT2020 and OpenAIRE (13 Nov, Bari, Italy).

EGI and the AARC project: the future of AAI

Peter Solagna gives an update on the progress so far

The EC-funded AARC project started in May 2015 as a collaboration among e-Infrastructures, NRENs, and other service providers, including various user communities and libraries. EGI is represented in the consortium by EGI.eu and several other partners of the EGI federation.

One of the goals of AARC is to deliver the design of an integrated and interoperable framework for Federated Authentication and Authorisation Infrastructures (AAI), which meet the needs of the Research Infrastructures and e-Infrastructures across Europe and beyond.

During the first months of the project, AARC members have been discussing requirements with the communities and the infrastructures, in terms of capabilities, blocking issues, and training requests.

The progress of this work is reported in:

- > a deliverable to *"Report on the identification of target groups and their requirements"*, published in August 2015, based on surveys and interviews of representative communities.
- > a deliverable to report on the *"Analysis of user community requirements"*, published in early October 2015
- > and the first Training Module, Federation 101, which will be constantly updated as feedback is received.

We can see patterns emerging from the feedback collected among the various stakeholders, namely in design and pilot implementation of federated AAls within the scientific communities and e-Infrastructures. An example is the deployment of identity provider proxy to enable multiple credentials sources. It seems that communities are choosing similar directions. These include also the EGI long tail of science platform developments and some of the communities involved in the EGI-Engage Competence Centres, for example. These commonalities are the starting point where interoperability can hopefully be achieved through harmonization of the solutions.

The goals of the AARC project are very well aligned with the EGI technical strategy. Support the increasing requirements for easy access and interoperability by implementing federated authentication and single sign on capabilities also on the EGI services. While AARC is looking into the policy framework and the architecture to support the use cases and expand the identity federation coverage, EGI-Engage project is focusing on piloting an integrated solution that will support the EGI major use cases. These use cases initially include enabling SSO and federated access to EGI resources and operational tools, and providing community

management capabilities to allow a multi-community/multi-provider federation as EGI to be fully productive. EGI activities will build on the AARC results, but also on the more than decennial experience of EGI in supporting federations and collaborations between user communities and service providers.

With six months of AARC behind us, the project is speeding towards the deployment of the first pilots focusing on attribute management, credential translation and guest identities provisioning. The project is also organizing trainings and workshops for the coming months. EGI will ensure that these outputs are available for our communities, and integrated with our AAI work.



EGI in AARC

About 1 FTE, spread across three tasks :

- > JRA1.1: Requirements gathering
- > SA1.2: Pilots of an attribute management framework
- > NA2: dissemination and communication

AARC project
<https://aarc-project.eu/>

Data Science Professional uncovered

How the EDISON project will contribute to define a career for Data Scientists

Data Science technologies are having an impact on how research is conducted, how scientists think, and how data is used and shared.

The effective use of Data Science technologies requires new skills and we need new people for this. Data scientists are experts capable to extract meaningful value from the data collected and manage the whole lifecycle of data, including supporting Scientific Data e-Infrastructures. The data scientists of the future have knowledge and skills in data mining and analytics, information visualisation and communication, as well as in statistics, engineering and computer science, and acquire experiences in the specific research or industry domain of their future work and specialization.

Currently there is no specific professional education for data science or generic big data technologies. The components of the data scientist skill set are offered, but are usually scattered across training programmes. There is not even a common approach as to how to effectively build professional level big data curricula.

We need to re-engineer what we have to create a model for data science education. This can be done by identifying existing building blocks and organise them into a coherent data science curricula, and address issues related to the size and complexity of data that are not in traditional education programmes.

The main challenges we foresee in this context are:

- (1) the analysis of requirements, which must be addressed with the data science industry and research partners,
- (2) the paradigm-shift for an alternative, data-centric view of traditional education programs, which must be addressed with the educators,
- (3) the integration of modern, data-specific challenges in this education program, which must be tackled by both educators and practitioners,
- (4) the domain specific data science education and availability of datasets and virtual learning environment.

The EDISON project has been established in order to support universities, research centers, industry and research infrastructure organisations to cope with the potential shortfall of Data Scientists, to define the framework of competences as well as the body of knowledge for this profession.

EDISON will define the data science profession in terms of a Competences Framework and Skills Profile (CF-DS), which will include common hard and soft skills required for successful data scientists.

The CF-DS will be complemented by a Data Science Body of Knowledge (DS-BoK) to compile the knowledge required by the profession.

We will perform an analysis of existing academic and professional courses and curricula,

including available books and other training resources. We will identify common conceptual elements and gaps amongst the present offering and we will create a certification scheme as a means to achieve recognition for this profession.

Authors

Andrea Manieri, *Engineering Ingegneria Informatica*
Engineering will lead the creation of the Body of Knowledge based on the CF-DS and a thorough analysis of the demand-side. It will also act as a proxy toward industrial stakeholders and specifically the Business Intelligence, Marine/ Maritime and Space markets.

Małgorzata Krakowian, *EGL.eu*
EGL.eu will manage EDISON's WP4 leading sustainability, business development and certification activities and contribute to the training WP3. EGL.eu will contribute to the body of knowledge for the certification scheme and act as a pilot client for implementation, training and certification.

PhenoMeNal: towards an e-Infrastructure for pheno- and genotyping data

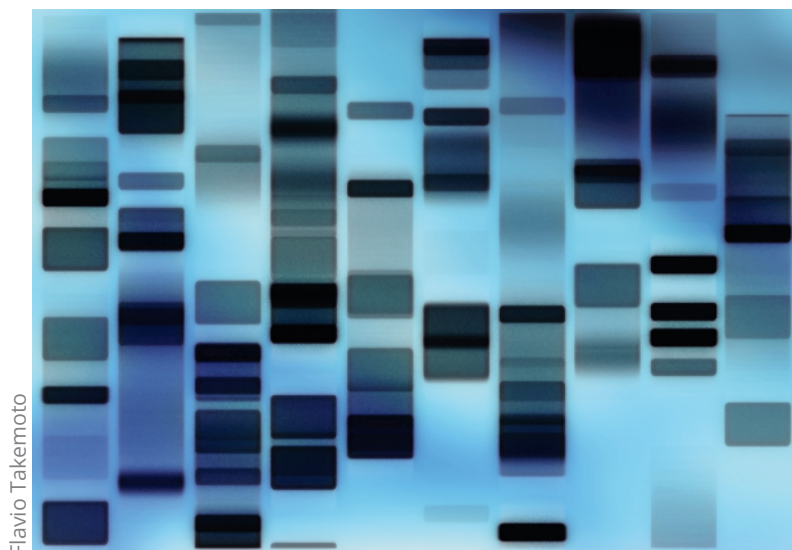
Namrata Kale and Christoph Steinbeck introduce a new project

Phenome and Metabolome aNalysis (PhenoMeNal) is a project set up to develop and deploy an integrated, secure, permanent, on-demand service-driven, privacy-compliant and sustainable e-Infrastructure for the processing, analysis and information-mining of the massive amount of medical molecular phenotyping and genotyping data generated by metabolomics applications entering research and clinic.

This e-Infrastructure will provide services to the European Biomedical Community enabling computation and analysis of molecular phenotype data to improve understanding of the causes and mechanisms underlying health, healthy ageing and diseases.

PhenoMeNal will address challenges arising from the large data volumes in molecular phenotyping by creating a federated and secure high performance e-Infrastructure to handle and analyse huge data sets.

Since no viable technological solution exists to transmit such large datasets between institutions in convenient time frames, PhenoMeNal will develop solutions that bring the compute to the data with virtualised compute engines that can be launched and run on the major available platforms. Wherever possible and advisable, PhenoMeNal will make the application available on national and European Grid and Cloud Infrastructures.



Teaming up with EGI

The EGI federated cloud capabilities offer a major opportunity for the PhenoMeNal project to deploy scientific pipelines in a vast number of academic clouds across the EU. This will allow PhenoMeNal to reach more prospective users and to ensure the sustainability of the project network.

In addition, the EGI's strategy to provide a secure and reliable infrastructure for the European Research Area is an added advantage for PhenoMeNal to establish its objective of privacy protection methods that allow working with highly sensitive human data.

The partnership

PhenoMeNal partners include lead partner European Molecular Biology Laboratory - European Bioinformatics Institute (EMBL-EBI, UK), Imperial College, the University of Birmingham and the University of Oxford in the UK, Leibniz-Institute of Plant Biochemistry in Germany, Universitat de

Barcelona in Spain, Consorzio Interuniversitario Risonanze Magnetiche di Metallo Proteine in Italy, Universiteit Leiden in Netherlands, Institute of Bioinformatics in Switzerland, Uppsala Universitet in Sweden, BBMRI-ERIC in Austria and Commissariat à l'énergie atomique et aux énergies alternatives (CEA), and Institut national de la recherche agronomique (INRA) in France. These partners have a vast experience in the development of tools and methods for acquiring, integrating and analysing huge datasets on metabolic phenotypes, genome and cross-omics data. In addition, the consortium has also successfully paved the way for this e-Infrastructure through their previous extensive standardisation efforts in COSMOS project (Coordination of Standards in MetabOlogicS) and metabolomeXchange – an international data aggregation and notification service for metabolomics.

New project: West-Life

Chris Morris writes about the goals of a project set up to support structural biology

West-Life is a H2020 project set up to pilot an infrastructure for storing and processing data to support the growing use of combined techniques in structural biology - the study of the structure and function of the macromolecules of life.

Structural biologists use many techniques including X-ray crystallography, Nuclear Magnetic Resonance (NMR), and electron microscopy. Typically, structural biologists are experts in one or more techniques. But as the field develops, they often find the need to use techniques with which they are less familiar. There are some technique-specific pipelines for data analysis and structure determination, but little is available in terms of automated pipelines to

handle integrated datasets, and integrated management of structural biology data from different techniques is lacking altogether.

This is where our new project comes in.

West-Life will integrate existing and future data management facilities. The result will provide users with an integrated overview of the experiments performed at different research infrastructures and links to the different data stores. It will extend existing facilities for processing this data. As processing is performed the system will automatically capture metadata, reflecting the history of the research. The effort will use existing metadata standards, and integrate with

them new domain-specific metadata terms.

West-Life will also develop application-level services specific to use cases in structural biology, enabling structural biologists to get the benefit of the generic services developed by EUDAT and EGI.



More information

West-Life will run for three years starting in November 2015.

<https://www.west-life.eu/>

Community Forum 2015: Data without boundaries

Sergio Andreozzi presents the data track (11 November, starting at 11:00)

The digitisation of science is a key driver of the exponential growth of research data across all scientific disciplines. At the same time, funding agencies and open science movements reinforce the openness principle. These two drivers impact the scientific process from technical to organisational and legal levels.

The track **Data without boundaries**, part of the EGI Community Forum, will address some of the key aspects related to the opening of data.

> The first session will focus on the legal aspects and interope-

rability of policies for the sharing and re-use of data. The goal is to identify issues and barriers across different disciplines (e.g. fishery, marine) and understand what approaches are possible to remove them.

> The second session will identify technical requirements from various types of stakeholders (e.g., public organisations, SMEs and industry) to understand what data would be useful to share and how to optimise access, through use cases driven by analysis specific scientific domains where data are generated.

> The third session will focus on metadata interoperability, by going through a number of initiatives and technologies, comparing approaches and identifying requirements.

Overall, this track will contribute to advancing the understanding on how EGI can better serve research communities and SMEs/Industries by offering an ecosystem of services and data that stimulate value creation across the whole chain from basic science to market exploitation.

<http://go.egi.eu/CF15wed>

Community Forum 2015: Preview

Highlights and information

The Community Forum will kick-off in Bari on Monday, 9 September with a day dedicated to EGI-Engage face-to-face meetings - <http://go.egi.eu/cf2015>

Registration for the conference is open until 4 November. After this date, registration will only be possible on site. <http://go.egi.eu/cf15-reg>

The full programme is available in the CF2015 Indico pages. <http://go.egi.eu/cf2015>

Social events

Welcome Party (Tuesday)

Drinks to welcome all the participants will be held in the Reception Hall of the Villa Romanazzi Carducci on Tuesday, November 10th soon after the end of the last parallel session.

Guided Tour and Cocktail Party (Thursday)

Since Bari is a city worth to be visited, the Local Organizing Committee has planned a guided tour to the most beautiful sites of the city. A nice walk around the ancient history and the modern peculiarity of the city, visiting some remarkable monuments and places that will never be forgotten.

The tour will end with a nice cocktail party in a wonderful restaurant, well known in the city for its cookery, where some Baresan Specialties will be served in a very easy and friendly atmosphere.

Keynote speakers

Augusto Burgeño, Head of Unit "e-Infrastructure" at the European Commission, will give a keynote talk on Friday 13 November entitled "*Opportunities and challenges of the e-Infrastructures and the new H2020 Work Programme 16-17*".

Tiziana Ferrari, Technical Director of EGI.eu, will speak at the opening plenary about the first six months of EGI and future plans

Co-locations

Co-locations will be held on Friday 13 November, after the closing plenary. Registration for the co-located events is free of charge, but it is mandatory.



- > Community workshop on Open Science Cloud
- > INDIGO-DataCloud project meetings (by invitation)
- > EDISON Expert Liaison Group meetings (by invitation)

Demonstrations and posters
Posters will be presented in the same area. List of posters <http://go.egi.eu/post>

Schedule of demonstrations <http://go.egi.eu/demo>

Citizen Science at EGI-Lifewatch CC: Orchids identification using pattern recognition tools based on neural networks over GPGPUs
Fernando Aguillar et al. Tue 10 Nov 9:00-13:00

High-throughput Processing of SEBAL in the EUBrazilCC FedCloud
Francisco Brasileiro et al. Tue 10 Nov 13:00-17:00

The EGI Training Infrastructure
Gergely Sipos et al. Wed 11 Nov 9:00-13:00

Deploying Cost-Efficient Virtual Elastic Clusters across Multi-Clouds
Amanda Calatrava et al. Wed 11 Nov 13:00-17:00

The new EGI platform for the long tail of science
Diego Scardaci et al. Thu 12 Nov 9:00-13:00