EGI Federation Service Strategy

2022–2024

About us

Advanced computing for research

EGI is a federation of computing and storage resource providers united by a mission of delivering advanced computing and data analytics services for research and innovation.

Our services include high-throughput and cloud computing, storage and data management, analytics, consultancy and support, training and co-development.

Scope and purpose

In 2019, the EGI Council approved the 5-year federation strategy 2020-2024. In the context of this work, the implementation of Strategic Goal 2 "Evolve the service offering to meet the needs of researchers" demands the development of a service strategy and related services roadmaps that will drive the research and development efforts of the EGI community and the evolution of the EGI service catalogue.

Today's EGI services deliver advanced computing services to support scientists, multinational projects and research infrastructures, and are organised in two catalogues. The internal catalogue delivers coordination and federation services to the data centres that are part of the EGI infrastructure, while the external catalogue addresses researchers, scientific communities, and innovators by delivering data-intensive computing, storage, data management, data analytics, trust and identity management, and training. The external catalogue services can be requested via the EGI Marketplace.

This document sets the priorities for the investigation of new or improved services considering the needs in data-intensive scientific computing gathered from research infrastructures, scientific collaborations, communities of practice and innovators. The creation of a Service Strategy is part of a larger effort to update the EGI Federation strategy and to define specific plans in selected areas. The EGI Service Strategy supports and complements day-by-day service portfolio management activities of the EGI Federation.

Target groups
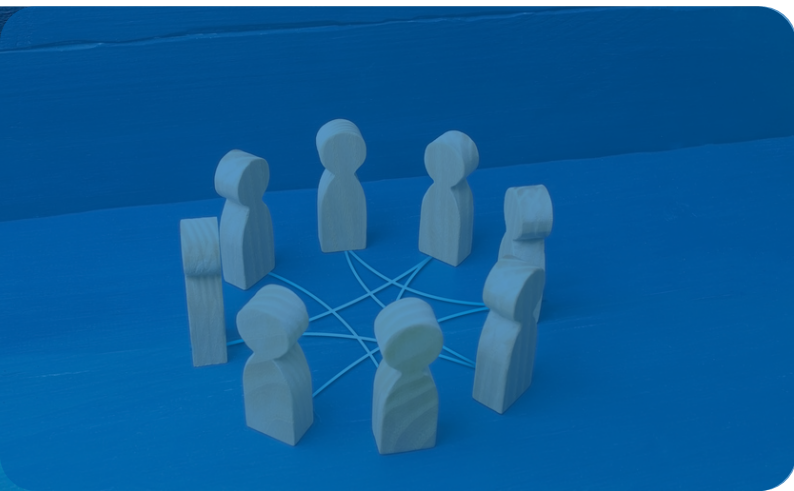
01 **Research communities and research infrastructures**

Research communities and research infrastructures push the frontiers of science, with instrumentations, massive data and computational requirements. They have an international nature, decades-long lifetime, and a strong need for shared e-infrastructures. We provide both services and expertise for community-specific solutions that help them serve their user base. The EGI Foundation acts as the main contact point with the head organisations, while the EGI Participants engage via their local contacts.

02 **International research projects and research collaborations**

International research projects and research collaborations are typically represented by H2020 EC-funded projects at the European scale. While such projects have limited lifetime they cover many niche domains and regions. To this target group EGI is offering IT services and technical support to run proof-of-concept and pilot setups to tackle the scientific communities' challenges, or, as Infrastructure provider, to provision and operate the communities' e-Infrastructures needs which can be used to analyse big data analytics and support advanced research.

03 **Small international groups and individual researchers**

Small international groups and individual researchers are usually less experienced in advanced computing or lack resources and support. The approach is to connect them to the most suitable service providers, through a common platform and engagement framework, or involve members of the EGI Federation based on the geographical distribution of the group or individual researcher, in order to reduce the level of engagement from the EGI Foundation.

04 **SMEs & industry**

SMEs/Industry generally interact with the EGI community in order to stimulate the knowledge transfer activities and exploit the main achievements produced in science and scientific innovation for business and society. The EGI Federation helps them to exploit services and resources for compute-/data-intensive research and innovation, support in reusing open research data sets, tools/applications, access to training/education courses and promote their services.

05 **Public administration**

Public administrations enable innovation in the policy making sector, removing European fragmentation, allowing cross-support and cross-collaboration and the use of secure compute and data-intensive services. Boosted by the Green Deal, and new projects for reducing the impact of Climate Change with Emergency Management (floods, wildfire, power outages), they are creating a new and very active Community. Health Care is also a fundamental area in which regional and local Authorities play a main role. They do not always own technical capabilities to develop solutions and need the support of Providers and EOSC initiative.

Strategic Objectives

This section describes five Strategic Objectives (SOs) for evolving EGI services. They are based on the knowledge of the user community needs that are being gathered through research and innovation projects, customer interviews and analysis of trends. Each defines an area for possible future work, and if approved, can be developed into more concrete goals and intended results.

They are described according to the following template:

Objective
What is the most important area to focus on making measurable progress?

Key results
How will we know we have made measurable progress?

Motivation for users
Why will our current/potential users be interested in this?

Motivation for the EGI Federation
Why should we prioritise it?

Use case examples
What are leading real-world examples?

Areas of work
What should we do to achieve it?

The following Five Strategic Objectives (SOs) are proposed and described in the following sections:
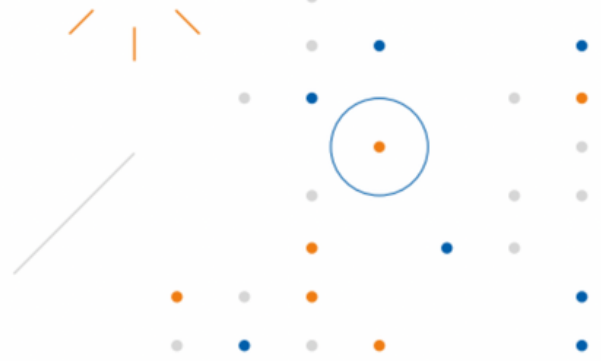SO1: Federated compute continuum
SO2: Federated Data Lakes and repositories
SO3: Data analytics and scientific tools including AI
SO4: Professional support and consultancy
SO5: Investigation of a trusted compute platform for sensitive data processing

SO1. A Federated Compute Continuum

Objective

Equip researchers with a federated hybrid compute infrastructure for convergent use of HTC, HPC, Cloud and Edge

Europe has significant data and computing resources that are of vital benefit to European research communities, however access and provisioning of capacity at national and European level need to be integrated to support heterogeneous workflows emerging from modern data-intensive and compute-intensive science from all domains involving modelling, simulation, image processing and other data-intensive processing tasks. Investments in compute capacity building and integration are paramount to address the data deluge in the coming decade.

Key results

- Existing and future user communities can run and scale up their workflows on the EGI Federation thanks to the availability of integrated HPC systems and accelerators from national compute infrastructures
- The distributed data processing model of EGI is further expanded towards the Edge where data is generated

Motivation for users

- Modern data-intensive and compute-intensive science from many key scientific collaborations involve modelling and simulation at very high resolution for prediction and inference workflows
- The complexity of the computing workflows and the data required by models, which can vary from Gigabytes to Petabytes of information per day to be processed, the ability to access heterogeneous computing architectures and the related data to be processed is translating into a leap in computational and data handling capacity needs

Motivation for the EGI Federation

- Integration of federated HTC and Cloud capacities. The integration has already been successfully demonstrated and adopted by research communities
- Expansion of existing federation solutions and policies to HPC and Edge resources
- Leveraging the national investments in HPC, HTC and Cloud and RI investments in Edge computing

Use case examples

- High Energy Physics experiments have long been using simulation tools for different tasks, ranging from the initial phases of detector design to the final analysis step to compare the experimental data to theoretical models. Gain in terms of computing time is essential in the High Luminosity LHC future experiments to enable large scale production of simulated data
- The sensitivity of gravitational wave interferometers is limited by noise; its reduction and subtraction are one of the most important and challenging activities in this field of research. Low-latency search pipelines to veto or de-noise the signal are required, sending out more reliable triggers to observatories for multi-messenger astronomy

Areas of work

- Harmonization of security and access policies
- Advancement of solutions for federated monitoring and accounting
- Evolution of the EGI trust and identity infrastructure
- Evolution of EGI supported software solutions for serverless computing solutions, workflow management, and efficient data transfer and mirroring

SO2. Federated Data Lakes and repositories

Objective

Equip European researchers with shared and federated high-quality research object data lakes for storing and analysis integrated with big data processing, real-time analytics, and machine learning solutions

Various EGI participants are engaged in the provisioning of national and thematic repositories for digital preservation of data, software, and other research objects. Data repository infrastructures are not yet federated at European level to offer a EU entry point for discovery, access and exploitation of data, bringing data next to computing and analytics services with related educational tools and user support. Besides this, value creation in data exploitation is generated by integrating data, domain-specific tools and the supporting compute and storage infrastructures. Practically, this means establishing a provisioning and support model for domain specific scientific applications integrated with EGI compute and storage infrastructure. Together, multiple data spaces should form a Data Science Commons, which aims to create a federation of research data, computing, applications, and other open science resources. This responds to the problem of scalable access to research data through a new data-centric approach to service provisioning that is complementary to the traditional data download model.

Areas of work

- Collaborate with research communities on data spaces, architectures, interoperability standards and business models
- Establish a data-centric business model and delivery channels in collaboration with research data providers
- Establish partnership agreements with research communities that play the role of data provider
- Consolidate federated data management solutions and data fusion solutions
- Collaborate with software providers to enrich data with adding value services, e.g. data imaging processing tools combined with archiving
- Repurpose existing solutions to facilitate access to big data for communities that don't have the federation capabilities offered by EGI
- Bring national and community initiatives together

Key results

- Data can be discovered, accessed, and exploited through European coordination
- National and thematic data repositories usage increases
- Data processing workflows are more efficient by avoiding the local management of large amounts of third-party data

Motivation for users

- Integration of data with appropriate processing tools
- Provision of core data resources from research communities and RIs in an integrated manner with computing
- Reduction of IT barriers in big data science resulting in higher productivity in research

Motivation for the EGI Federation

- Promotion of MS funded national infrastructures for data deposition from emerging communities and processing and contribution to the EU priorities such as the EC data strategy and the EOSC SRIA
- Leveraging several national initiatives that are setting up national repositories for research data
- Provision of a federated trust framework where data can be shared and exploited among authorized members of a given scientific community thanks to EGI federation security management tools and policies
- Further exploitation of existing technical solutions to also provide secure access to public data as a service

Use case examples

- Climate modelling requires the management of an increasing amount of data near computing and require an easier access to climate dataset analytics, as well as AI-based analysis of very large ensembles of datasets
- Efficient processing of imaging data from environmental sensors requires co-location of large imaging datasets and computing integrated with AI-powered analytics tools

SO3. Data analytics and scientific tools including AI

Objective

Equip researchers with scientific applications, data analytics tools and AI applications that provide an integrated point of access to European core data resources and computing

Data-intensive science is enabled by thematic services that provide an easy-to-use software environment customized to the needs of scientific communities. These virtual research environments are a key access point to distributed computing, the EGI Federation can play an active role in collaborating with research communities to provide a hosting environment in the EGI federated cloud, promote them and support training and technical user support to expand their adoption. In addition, AI is increasingly adopted in research. The EGI Federation can provide AI services on demand in particular for training of models with distributed cross-border data.

Key results

- Users can choose from a large portfolio of scientific applications integrated with the EGI Federation and research data
- Users can run Jupyter notebooks on-demand with access to data and tools for AI and deep learning and train their models on GPUs
- Users can share models and exploit them in the cloud

Motivation for users

- Support with cloud-based application hosting of thematic services in the cloud
- Access to simulation tools to understand complex and multidisciplinary physical phenomena in their research project.
- Acceleration of the research process and increase of its efficiency by contributing to the development and use of ready-to-use tools.
- Access AI/ML applications and receive support for their development and operations

Motivation for the EGI Federation

- Differentiation and extension of the portfolio of computing services with a bespoke solution and adding value to innovative products not available commercially on the market
- A remarkable increase in the importance of AI/ML technologies to research and science, economy and society. More and more researchers are turning to machine-learning models to develop accurate, fast predicting models on a wide range of subjects

Use case examples

- The EGI Federation has been supporting the integration, operations, and user training of the WeNMR portfolio for molecular simulation tools reaching an unprecedented scale of adoption with more than 26,000 users at the time of writing
- Data infrastructure in marine biology and in other environmental disciplines require the integration of multidisciplinary data, analytical tools and computing facilities (e.g. the BlueCloud demonstrators)
- LETHE project develops a personalized prediction and intervention model for early detection and reduction of risk factors causing dementia, based on AI and distributed Machine Learning. The project uses EGI FedCloud services for sensitive data storing, sharing and processing
- AI4PublicPolicy is a joint effort of policymakers and Cloud/AI experts to unveil AI's potential for automated, transparent and citizen centric development of public policies. To this end, the project will deliver, validate, demonstrate and promote a cloud platform based on EGI Cloud services and EOSC offers

Areas of work

- Select and establish collaborations with motivated and reliable RIs/user communities willing to co-develop with EGI their scientific applications, and provide funding and IT resources for their operations and promotion
- Collaborate with members offering a wide library of AI/ML analytics for researchers (datasets, models, algorithms)
- Provide adequate resources to deploy AI/ML algorithms and platforms (storage, computing, CPU, GPGPU etc.) and the necessary training
- Create a team of experts (a Competence Centre) on AI/ML in EGI

SO4. Professional support and consultancy

Objective

Users receive professional consultancy and technical support for their complex digital needs

Data-intensive scientific collaborations rely on very complex software solutions that need to be tailored to the specific needs of the community. The delivery of a production-quality software stack can require various years of engineering and development that need to be facilitated by IT experts. Specialized support and consultancy services are necessary to scale up the adoption of the EGI service catalogues. These services will answer to the increasing request of coordination, problem solving and technical orientation in the wide scenario of projects, solutions, partners, and customers, along with applicable business models.

Motivation for users

- Modern data-intensive and compute-intensive science from many key scientific collaborations involve modelling and simulation at very high resolution for prediction and inference workflows
- The complexity of the computing workflows and the data required by models, which can vary from Gigabytes to Petabytes of information per day to be processed, the ability to access heterogeneous computing architectures and the related data to be processed is translating into a leap in computational and data handling capacity needs

Use case examples

- All scientific communities, Digital Innovation Hub pilots and the EGI-ACE data spaces

Key results

- EGI participants actively participate in technical support to user communities with their experts

Motivation for the EGI Federation

- Scale up with the increasingly high demand tapping into national and research institute-level expertise. The current number of technical support experts is too limited in the EGI Federation, and it probably lacks engagement with existing competency centres
- Extension of the existing user base by reducing barriers in accessing the necessary facilities and IT services for data-intensive solutions by providing the necessary expertise
- Ability to address the complexity of the digital needs of research communities and the requirement of long-term support to customize solutions to existing IT systems
- Ability to address the lack of skills and shorten the time to adoption of existing solutions

Areas of work

- Increase funded participation in user community projects of EGI federation members
- Develop a network of experts providing expert advice and create a sustainable model that is not dependent on projects - understand status of national support structures that can be leveraged by the EGI Foundation central team
- Improve quality and quantity of training resources
- Develop the business models (and potential pricing scheme)
- Collaborate with other infrastructures and with research infrastructures on a common training programme, leveraging EOSC opportunities. Collaborate with research communities to increase the reach to scientific communities
- Participate in EOSC training initiatives and projects

SO5. Investigation of a trusted compute platform for sensitive data processing

Objective

Investigate the cross-border demand for sensitive data processing services, existing solutions, EGI federation data centre capabilities

Based on experience in projects relevant to Health Science like HealthyCloud and LETHE, EGI is going to mature knowledge of needs and expertise to manage data privacy in distributed computing. A subset of large data centres of the federation has the potential to act as Sensitive Data Management infrastructure offering secure user environments for secure data processing with controlled Access Services, Secure Storage and VMs.

Motivation for users

- The growing need for secure processing of sensitive data on HPC systems is challenged by high-performance computing clusters operating in shared and batch modes
- Shared computing facilities are very popular due to their accessibility and cost-effectiveness. Typically, clusters operate in a shared and batch mode, which means that many users can run jobs simultaneously on separate sets of compute nodes in the cluster. The shared nature of these systems, however, makes it challenging to impose security requirements for processing highly sensitive or confidential data
- High demand for computing facilities with a tightly managed security level due to increasing availability of sensitive and confidential data, and data protection laws, such as GDPR in Europe, and HIPAA in the US coming into force
- Fine-grained control over data import and export, in combination with auditing and logging of data access and export, will meet the data protection requirements. These are needed in studies where researchers use data made available by third parties, and the data owners will impose restrictions on the export of data and results. These facilities are typically not present on current computing systems

Use case examples

- Projects from Social Sciences and Humanities require access to governmental data about citizens and national solutions are available to share such confidential data with selected users in a controlled manner, however no mechanisms exist for cross-border access to such data allowing for a broader exploitation
- Clinical research benefits from the sharing of data from large cohorts of patients to increase accuracy of studies and treat rare diseases, such data is privacy sensitive and usually only accessible to on-site institutions.

Key results

- Awareness of cross-border sensitive data management requirements
- Awareness of distributed sensitive data management solutions and capabilities of data centres participating in the EGI federation

Motivation for the EGI Federation

- Ability to meet the unmet demand from user communities needing to manage secure data across borders and institutions because of the lack of federated solutions for secure data management in the HTC and Cloud federation. In particular, we lag behind in support of social sciences and health science
- Lack of IT centres offering the necessary ISO 27000 certification. The certification is requested by several communities dealing with secure data
- Enable multiple players acting over different sectors and countries to cooperate and access shared resources, services and data according to their regulatory context through the cross-border pan-European federation computing platform model
- Extension of the capability to access the sensitive data in protected environments and with encrypted protocols for data in transit and at rest will support the rapid development of services for the many user segments.
- Improvement of the trust of the users through a correct awareness and training in ISO27001

Areas of work

- Scouting of production ready solutions that can scale up for cross-border sensitive data processing and repurposing of existing solutions across research infrastructures
- Innovation in sensitive data management, leveraging solutions that have been implemented at local and national level, for example supporting the 'data visiting' paradigm rather than the 'data sharing' model
- Exploration of approaches such as secure VMs for data transfer
- Collaborations with support activities to the health and social science communities

EGI Federation – Service Strategy
2022-2024