Secure Federated Data-Analysis Capability for the European Research Area

Tackle society's grand challenges by providing a powerful, secure, efficient and scalable research infrastructure and associated support services for leading-edge data analytics

Motivational Context

The European Research Area (ERA) will need to support researchers from diverse scientific disciplines taking approaches to data analysis. These will need to work seamlessly together in a distributed multi-disciplinary research collaborations that cross national and intellectual borders to tackle society's grand challenges. For the ERA to successfully increase the ability of Europe to produce 'excellent science', which delivers exploitable innovations and new growth, Europe's researchers will need easy to use integrated services that provide access to high capacity and high quality computing and storage resources, wherever the resources and the researcher are located.

Over the last decade, the European Grid Infrastructure (EGI) has built a distributed computing and data infrastructure to support multi-disciplinary science. This infrastructure has since delivered an unprecedented data analysis capability to over 21000 researchers, including the High Energy Physics community in their successful search for the Higgs Boson particle using the petabytes of data generated by the Large Hadron Collider at CERN.

The EGI Collaboration (through its governing body the EGI Council) is composed of representatives from the National Grid Initiatives (NGIs) and supported research communities, is now proposing an ambitious investment in *open distributed computing and data infrastructures* that will build on the existing European and national investments to:

- **Provide Enabling Services to Researchers** by adopting a defined service portfolio and a user-centric approach to its development that will expand EGI's current service offering from EGI.eu and affiliated NGIs to retain its current research communities and attract new research communities.
- **Operate an Unprecedented European Capability for High Throughput Data Analysis** that expands EGI's current federated cloud infrastructure to 10M computing cores and 1Exabyte of storage by 2020. EGI will build on its current collaborative resource allocation model based on the resources coming from the NGIs and other organisations to also support peer-reviewed or pay-for-use access for researchers undertaking excellent science.
- **Provide Flexible Virtual Research Environments** that simplify access to EGI's resources and accelerate the ability of researchers to undertake excellent science by leveraging the expertise and connections of the NGIs to introducing technical innovations into production across Europe.
- **Identify and Develop EGI's Human Capital** to upskill the research communities supported by EGI in establishing within NGIs national centres of excellence that can transfer skills from within EGI to tomorrow's data scientists.

Collectively, these activities stimulate technology innovations and result in research innovations through rapid and effective use of high-throughput data analytics on a high-capacity and integrated European infrastructure by research communities in the ERA.

Activity Overview

Providing Enabling Services to Researchers

- **Issue**: Different researchers and research communities all have varying requirements on the e-Infrastructure services and resources they need to meet their data analysis needs. European e-Infrastructure providers (i.e. EGI, DANTE and PRACE) are increasingly being challenged to meet the needs of all researchers through integrated services.

- **Current Status**: EGI has over the last two years been defining its technical service portfolio provided centrally through EGI.eu and its affiliated resource providers to clearly define the capabilities it provides to the different research personas which respectively represent the needs of large global research infrastructures (e.g. WLCG, ELIXIR), small/medium research collaborations (e.g. WeNMR, DRIHM) and enterprises, and finally the individual researchers in the long-tail (e.g. supported by the European Science Foundation).
- **Future Plans**: With further funding EGI will continue to evolve its solutions portfolios by establishing user boards to manage the definition and development of requirements, roadmaps, whitepapers, metrics, training and promotional material. Each solution will have a dedicated manager providing the 'voice of the user' who will drive change through the introduction into production of technical innovations with relevant service and technology providers.
- **Impact**: Early in Horizon 2020, EGI will have established an integrated service portfolio from European and National service providers (e.g. NGIs) and have adopted a user-centric development model where user requirements are met by the best service providers from the public or private sectors. As a result, EGI will be able to consistently improve the services to its current consumers and target new research communities through targeted developments. By the end of Horizon 2020, the uptake of EGI's services will have diversified and the usage increased alongside new co-developed technical innovations and related service capabilities (by bringing research communities, service providers and technology experts together) will have emerged and been brought into production to meet the new research challenges in the open compute and data infrastructure community.

Operate an Unprecedented European Capability for High Throughput Data Analysis

- **Issue**: For the ERA to undertake excellent science, research communities need access to services that meet their different requirements supported by an e-Infrastructure capacity (including storage, networking, computing, …) that allows researchers working in large or small collaborations, or individually to undertake their data intensive science. Increasingly, the needs of individual research collaborations are exceeding the capacity of a single resource centre or country to meet their needs thereby accelerating the move to distributed resource provisioning models.
- **Current Status**: For over a decade, EGI has been operating a reliable secure federated infrastructure composed of computing and storage resources contributed by NGIs, EIROs and other organisations consisting in June 2013 of 361K computing cores and over 400PB of disk and tape storage. This production infrastructure operates 24/7 and through its redundant and distributed architecture delivers 100% availability to the major research communities that depend on it for their data analysis needs. Access to this capacity is currently driven from the bottom up by the research collaborations and their associated resource providers from within the NGIs.
- **Future Plans**: EGI.eu will continue to operate a set of core services while providing the operational and technical coordination needed to reliably and uniformly federated distributed resources from across EGI's resource centres. In a federated infrastructure, national funding bodies will continue to fund and manage access to the resources coming from the NGIs. With further European funding, EGI will build on its existing resource allocation mechanisms to establish a European level peer-review process that will enable research collaborations undertaking 'excellent science' to access pooled resources. For those research collaborations able to pay-for-use resources, EGI.eu will coordinate access to both public sector resources able to provide payment based access and integrated private sector resources.
- **Impact:** By integrating new resources into EGI's current federated cloud infrastructure, by 2020 EGI expects to operate a cloud infrastructure composed of public sector resources coming from resource providers within the NGIs and other organisations comprising over 10M computing cores and 1Exabyte of storage.. These resources will be available to the ERA accessible through peer review or through pay-for-use models to undertake excellent science. The federated model will allow researchers to scale out their data analysis capability beyond the capacity of a single resource centre and to access these trans-national resources securely, flexibly and reliably at unprecedented scale.

Provide Flexible Virtual Research Environments

- **Issue**: The data analysis environment needed by researchers to undertake their excellent science will depend on a range of services coming from the NGIs and other service providers from the public and commercial sectors. Assembling the unique easy to use virtual research environment needed by an individual researcher requires a mixture of generic services that will be used by all research communities, to bespoke services used by just a single research community.
- **Current Status:** Members of the EGI community have been working with their NGIs and affiliated technology experts, service providers and research communities in Europe to establish the virtual research environments needed by research collaborations to effectively exploit the distributed computing and storage resources. These are presented to researchers as web portals, or as mobile or desktop applications customised to meet their specific research needs. The current software services provided by EGI are currently being used by over 21,000 researchers as the foundation of their own domain specific virtual research environments.
- **Future Plans**: To deliver high-quality solutions that will reliably scale to the environments presented by a distributed computing and data infrastructure, further European and national funds are needed to co-develop software through a user-centric model. EGI will facilitate bringing together service providers and technologists within the NGIs and researchers with challenging requirements to co-develop new services that will increase the capability of researchers to undertake excellent science.
- **Impact**: Easy to use virtual research environments tuned to the needs of specific research communities will lower the barriers to those researchers in using EGI's open computing and data infrastructure to support their data analysis needs. EGI will facilitate the development of these services by bringing together research communities, technology experts and service providers to co-develop the services that researchers need. EGI can provide a mechanism to sustainably operate those services that enter widespread use within the ERA and evolve the service offering to meet the needs of those that use them during Horizon 2020 and beyond.

Identify and Develop Our Human Capital

- **Issue**: EGI depends on human networks and the capital within it that spans NGIs, technology experts and research disciplines to operate the infrastructure, provide outreach to existing and new research communities and the technologies upon which the infrastructure is based. These human networks currently have different levels of maturity, connectivity and skill base across Europe and these all need to be improved.
- **Current Status**: In recent years, EGI has developed and grown its human networks from just the NGI operations centres to also include other NGI non-operational activities (e.g. policy, communications, promotions, events, technical outreach, etc.), and leading young researchers who are already using EGI. These human networks are coordinated by EGI.eu centrally to provide communication channels and the human capital within these networks is currently being developed through webinars and F2F meetings. Social media tools are used to strengthen communication within and between these different human networks.
- **Future Plans**: With funding EGI will continue to develop and grow the human networks based in the member states through the NGIs, research disciplines through the EGI Champions and experts from the technology teams that produce the software used by EGI, and developing other human networks as they are recognised.
- **Impact:** The human capital within these networks, which resides within and outside the NGIs, will be developed through coordinated training and strengthened through relevant events. The expertise within these networks will be available to researchers through NGI and regional virtual centres of excellence which will provide a focal point for dedicated training resources, application development and consultancy to researchers across all disciplines in their adoption of e-Infrastructure services, thereby increasing and dispersing the expertise relating to the use of e-Infrastructures across Europe.

Impact

For the researchers and research communities within the ERA

- **Supporting Excellent Science**: Researchers will have access to their local and affiliated NGI resources through their existing access mechanisms. EGI provides access to additional resources throughout Europe for work leading to excellent science that can tackle society's challenges. This is currently achieved through the researcher's own research collaborations but in the future additional resources allocated by peer-review, or by pay-for-use using integrated commercial or public sector resources could also be made available. The peer-review allocation process of EGI's resources is currently being prototyped and the integration with commercial cloud resources is being explored through the HelixNebula initiative.
- **Making every Researcher Digital**: The new research approaches being driven by the data deluge require new technologies and new skills for researcher's to fully exploit this new research paradigm. EGI provides European coordination of human capital distributed across Europe that is marshalled nationally by its affiliated NGIs, that can be used to co-develop new technology with researchers, and to transfer these technology skills to the applied research community through national competency centres. EGI's Champions and NGI contact points (in operations, user support, policy, communications, promotion, etc.) will be developed so that their skills and experience can be transferred (through training, application development and consultancy) to other communities.
- **A Scalable Computational and Storage Research Infrastructure**: Researchers and research communities undertaking data driven science will find their research capability limited by the infrastructure services and capacity that they have available locally for their data analysis. EGI provides a service infrastructure using integrated NGIs that allows researchers to transparently and uniformly access and use resources, data and knowledge across Europe wherever they are located, from both the public and commercial sectors. Further investment will increase the automation and reduce the central (EGI.eu) and national (NGI) operating costs of EGI while enabling cloud and other resources to be fully integrated into Europe's production infrastructure.

For NGIs and other organisations participating in EGI.eu

- **Growing and developing a national e-Infrastructure**: The capacity and capability of national e-Infrastructure (compute, storage, networking, training, outreach, education, …) varies greatly. EGI through EGI.eu provides a focus point for integration and collaboration at a European level of comparable European activities (PRACE, GEANT, EUDAT, HelixNebula) and that can complement closer integration and development taking place at a national level within NGIs. Such integration can provide a single point for researchers to access e-Infrastructure nationally and EGI will explore with other stakeholders how a single point of access and integration can be provided for researchers at a European level.
- **NGI Virtual Centres of Excellence**: New skills will be needed by a member states researchers' to access and effectively use the new technologies needed to exploit the data emerging from the excellent science being undertaken nationally as part of the ERA. NGIs are able to help transfer these new technologies and techniques (e.g. training, application development and consultancy) from the EGI community to their national research communities through NGI virtual centres of excellence. Such activity will develop the human capital across Europe within EGI and export these skills to the research communities within the ERA.

For the European Commission and European citizens

- **European Coordination and Governance**: EGI brings together over 35 organisations representing member states and research communities as participants in EGI.eu (an independent not-for-profit Dutch foundation), in a common mission to provide an open computing and data infrastructure across Europe. In addition to providing technical and political coordination and governance EGI.eu federates national activities into a European whole allowing researchers of all disciplines to have transnational access to resources through their ability to undertake excellent science regardless of the researcher's location, discipline and nationality. Proposals to explore a Digital Research Infrastructure European Research Infrastructure Collaboration (DRI ERIC) were recently endorsed by the EGI-InSPIRE project review and seen as driving greater European and

national integration of e-Infrastructures that could both improve service quality from a user perspective and sustainability.

- **A Distributed Open Computing and Data Infrastructure for Europe**: The provision of computing and data capacity in Europe resides in NGIs and other organisations, yet increasingly needs to be delivered at a European level to meet the needs of European Research Infrastructures (e.g. ESFRIs) and the global collaborations in which they participate to deliver excellent science. As a European coordination body, EGI.eu provides the means to integrate uniformly the new technical computing, storage and data innovations needed to provision these resources across Europe for the benefit of the ERA and its researchers thereby preventing the emergence of an e-Infrastructure digital divide and to represent Europe worldwide in providing e-Infrastructure.
- **High Utilisation of Public Sector Investments**: Member states continue to make significant independent national investments in their own physical e-Infrastructure resources (e.g. clouds, grids, desktops, data, networks, HPC), yet increasingly their researchers work in European wide collaborations. EGI's production infrastructure allows authorised researchers to use capacity from across Europe to meet their data analysis needs that could be delivered through EGI using the existing collaborative resource allocation model or additionally by a peer-reviewed allocations process or by an internal market with public and commercial sector resource supporting pay-for-use. This would ensure that NGIs fully utilise their national infrastructure and could by supporting pay-for-use on excess capacity also see a financial return on these investments if so desired.

The European Landscape

This paper is a refinement of EGI's strategic vision 'Seeking New Horizon – EGI's Role in 2020'[1] that was published in June 2011. The paper helped focus discussions within EGI and the role of EGI within the broader e-Infrastructure Landscape. This is reflected in the latest e-Infrastructure Reflections Group (e-IRG) roadmap[2] and white paper[3] which provides a vision for 2020 where Europe has:

> *a single "e-Infrastructure Commons" for knowledge, innovation and science, as a living ecosystem, which is open and accessible and continuously adapts to the changing requirements of research.*

A key aspect to such an e-Infrastructure Commons is the uniform reliable delivery of services designed and continuously evolved to meet the needs of the research community. EGI's focus on working with research communities to continuously develop and *provide enabling services to researchers* and then deliver these services by *operating a high throughput data analysis capability* contributes directly to the delivery of excellent science in an e-Infrastructure Commons. The e-Infrastructure Commons will need to be established to meet the needs of the European Research Area within the European Commission's Horizon 2020[4] program. EGI's is committed to *provide flexible virtual research environments* co-developed in partnership with the research community to bring new innovations into production and to *identify and develop EGI's human capital* with the data science skills needed to exploit these new data analysis tools within the research community.

Collectively, the identified activities will with investment from European funding bodies, stimulate technology innovations and result in research innovations through rapid and effective use of high-throughput data analytics on a high-capacity and integrated European infrastructure aligned to the needs of research communities in the ERA.

[1] https://documents.egi.eu/document/1098
[2] http://www.e-irg.eu/images/stories/publ/e-irg_roadmap_2012-final.pdf
[3] http://www.e-irg.eu/images/stories/dissemination/e-irg_white_paper_2013_short_version.pdf
[4] http://ec.europa.eu/research/horizon2020/index_en.cfm?pg=excellent-science