EGI Position Paper for Data Services

Authors: Gergely Sipos, Peter Solagna, Salvatore Pinto, Tiziana Ferrari, David Wallom
Date: 2/Dec/2013
Document status: draft
Document link: https://documents.egi.eu/document/2038

1 Abstract

This position paper has been prepared as input for the 'EGI Towards Horizon 2020 Workshop' organised in Amsterdam in December 2013. The paper aims to take the first steps towards defining an EGI strategy in the data domain for the 2014-2020 period. The strategy should build on EGI's existing capabilities in the grid and cloud computing area that already support data-driven science in various scientific disciplines. The strategy should consider and respond to the new opportunities and needs that emerge from EGI's partner e-infrastructures, from various technology projects and from scientific communities of the European Research Area. The paper provides a summary of some emerging, representative scientific use cases that require data services currently unavailable from EGI. These use cases have been collected by members of EGI.eu during November 2013. The paper explores the possible responses that EGI could give to the new use cases, and proposes specific responses for most of the cases. The paper also identifies a few cross-cutting questions that the EGI community should answer before we can proceed with defining a data services strategy for Horizon-2020.

The next version of the paper will be prepared after the EGI Towards Horizon 2020 workshop based on the feedback received at the event, and in email. Feedback in email can be sent to the corresponding author: Gergely.Sipos@egi.eu.

2 Data use cases and requirements

This section is not meant to be a comprehensive description of all emerging use cases, but to highlight the most representative use cases of the existing and expected future beneficiaries of EGI. The use cases have been captured by members of the EGI community during 2013.

2.1 Scalable, personal storage

Scientific communities want to offer domain-specific e-laboratories for their members/partners. These e-laboratories are local or remote (cloud) installations of virtual laboratory software that can be customised with services, applications and data according to the collaborators' needs. E-laboratories should enable the importing of data from different 3rd party sources for integration, curation, processing and visualisation. This requires scalable, personal storage spaces that can be attached to e-laboratories on demand. Current EGI storages (SRM) provide space that is shared by all members of a VO.

Source: Lifewatch, BioVeL

Potential solution(s): dCache has a related development in its roadmap?

2.2 Metadata discovery

The community stores data and metadata together (for example in files or streams). Members of the community would like to perform searches on the data using queries on the metadata. They require a system that can discover metadata in the data sources, can catalogue the data based on the discovered metadata, and can perform queries on data based on metadata. The solution should be a framework to which data sources and metadata discovery algorithms/services can be connected to, and that can scale up to 10th of TBs of datasets. The framework should provide data discovery facilities through web interfaces. The framework should also enable the post-processing of the discovered data on EGI computing services. (Integration with HTC grid and/or cloud services)

Source: EISCAT_3D

Potential solution(s): Open Source Geospatial Catalogue hosted on the EGI Federated Cloud (Development of a Proof of Concept setup is ongoing in the ENVRI project) and Distributed Competence Centre to develop domain-specific metadata extractors for the various user communities.

2.3 Long term preservation

The community requires services for long term data preservation. Data preservation develops on several levels: bit preservation, data preservation, metadata preservation and software preservation. Almost all these activities require the execution of computational tasks to convert the data (or the metadata), to test the software and the framework that analyse the data, and to process the data in order to add annotations and metadata.

Source: High Energy Physics (HEP), Digital Cultural Heritage Preservation (DCH-RP), EISCAT_3D, EMSO, EPOS

Potential solution(s): data curation tools and frameworks, virtualized solutions for software testing,

(eg. Zenodo for small datasets and software; EGI Applications Database for software; PURL for large datasets)

2.4 Services for citizen scientists

Providing scalable services for scientific projects to store and curate data submitted by citizen scientists to them. Such services and infrastructure include provisions for integrating existing, local epistemologies (i.e. knowledge and data collections accumulated by citizens with specific or particular interest in a scientific discipline) into the public body of scientific knowledge, provisions for IPR management, and cost recovery for citizen scientists and research institutes that host civic epistemologies.

Source: DRIHM, DCH-RP

Potential solution(s): EUDAT Simple Storage for DRIHM?

2.5 Data with access restrictions

Providing storage and processing services for data that have access restrictions (e.g. because of some ethical, legal or societal reason). For example

- The system should guarantee that no one besides the producer of the data has access to the data. OR
- The system should guarantee that the data cannot be downloaded, only processed by other users (and these users can access the derived data).

A particular implementation of the concept is EBI's Embassy cloud. Embassy cloud (will?) enable users to upload Virtual Machines into the Embassy cloud where the data is hosted. The Virtual Machines can mount, then process the data. The user can access the processed/derived data, but not necessarily the original data.

Source: Life sciences, Economy?

Potential solution(s): Hosting confidential data in the EGI Federated Cloud, and allow access only through certified Virtual Machine images?

2.6 Data preservation from science gateways

EGI could support scientific gateways to transfer users' computational results from the gateways to repositories where the data can be preserved for long term after being properly indexed with metadata for later reuse. It would no longer be the responsibility of the scientific users to save and share data, but this would be offered directly by the gateways. In such a scenario, it will be important to involve the scientific users in a user-friendly manner, i.e. allow them to put an embargo on the data and/or delete those when irrelevant. Having the data properly stored and preserved adds additional value to them by providing a variety of post-processing and analysis tools that can work on the shared data. This should make such federated open data repositories attractive to both end users and software developers.

Source: WeNMR (structural biology)

Potential solution(s): Development an API for federated gateways on top of long term preservation services?

2.7 Open Data services

Provide storage for data generated by projects that want to make the data public and indexed in OpenAIRE. These projects require not only storage, but also the high level software environments, APIs and support services that provide a complete solution.

Source: EC H2020 documents

Potential solution(s): ?

2.8 Close compute and data in the cloud

Provide a federation of cloud storage and cloud compute services to seamlessly process data without the need of moving it around (large data transfers). In this scenario, Virtual Machine Appliances are uploaded by the user and instantiated in an IaaS Cloud as close as possible to the site which stores the data. The data access can be discovered from a single entry point and accessed via the cloud compute service.

Source: BioVeL, ESA SSEP, ELIXIR

Potential solution: Open Search solutions?

3 EGI strategy

3.1 Possible options

EGI provides a set of platforms and support teams that can implement and support data-intensive scientific use cases. The services that EGI provide in this respect are:

1. **A scalable file storage** in the form of a federation of grid services exposing common data access interfaces (SRM). The file storage is enriched with file catalogues and with metadata catalogues (LFC, AMGA). The federated file storage is part of the EGI HTC solution and it is interfaced with grid computing services.
2. **A scalable block storage** in the form of a federation of cloud sites exposing Cloud Data Management Interfaces (CDMI). The block storage is part of the EGI Federated Cloud platform and can be mounted by Virtualised Appliances running within the same cloud.
3. **A set of core services** that enable the federation of new types of data services into EGI's production infrastructure. These core services provide common authentication, accounting, monitoring, helpdesk and information system for the federation.

Given the above set of four types of services how can EGI respond to each of the use case requirement that are described in Section 2? The possibilities are the following (or a combination of these):

1. **Extend grid storage:** EGI can choose to extend the existing EGI scalable file storage with new capabilities that would make it capable of addressing new requirements. This means the further development of existing grid storage services, or interfacing higher level tools with these services to get a more complex service setup that can support the new use cases. The development work would require experts of the current services, and representatives of the scientific community to drive the development with detailed requirements.

2. **Extend the cloud:** Extend the existing EGI Federated Cloud platform with new capabilities that would make it capable of addressing new requirements. This means the integration of new storage/data-related services into the EGI Federated Cloud portfolio besides the current OCCI and CDMI offerings. The task would require technology experts who can identify and federate new services into the EGI Federated Cloud. These should use open standards and interfaces so any cloud site of the federation can deploy the new services. Representatives

from scientific communities are needed to drive the technology selection and developments with detailed requirements.

3. **New service in the cloud:** EGI could bring new types of services into its production infrastructure by deploying those in the form of Virtualised Appliances on the EGI Federated Cloud. The task would require cloud experts who can turn external services into virtualised services that are capable of running on the EGI Federated Cloud through its OCCI and CDMI interfaces. Representatives from scientific communities are needed to drive the developments with detailed requirements.

4. **Federate new services:** EGI can federate new types of data service into its production infrastructure by connecting these services to the elements of the EGI core platform. The new services should address the new requirement and can be operated by the EGI resource centres for the scientific communities. This task would require a technology integration project that identifies and takes software from external developers, for example from EUDAT, Pandata, and interface these services with the EGI authentication, accounting, monitoring, information system and helpdesk services. The project would require technology experts who can interface software with the EGI core services, Resource Providers who could operate the services, and representatives from scientific communities who could help identify external software to be federated and could drive the integration activity with detailed requirements.

5. **Act as a technology provider:** EGI could choose to act as the integrator of software that is developed inside or outside of EGI, and package this software for scientific communities who want to operate data services for themselves. EGI would not operate services based on the software, and act 'only' as a technology provider for the scientific community. Such a project would require software integrators and testers, optionally software developers (if the development is performed in EGI), and representatives of the scientific community who would operate and provide services for their community based on the software.

6. **Do nothing**: EGI can choose not to respond to the requirement, for example because we expect solution to come from some other e-infrastructure, we do not see sufficiently large user base for the use case, or because we cannot offer a sustainable solution.

3.2 Proposed responses to the use cases

Given the use cases in Section 2, and the possible EGI responses in Section 3 this section describes a provides a proposal for the EGI community about how to respond to each of the use cases. This table should be updated and completed based on the outcome of the EGI Towards Horizon 2020 workshop.

Use case	Which strategy should EGI follow to support this use case?	What will EGI do next to support this use case in Horizon-2020?
Scalable,	3. New service in the cloud:	

personal storage	Identify and bring into EGI an external solution that builds on CDMI and could be hosted as an SaaS in the EGI Federated Cloud.	
Metadata discovery	3. New service in the cloud: Complete the implementation of the Open Source Geospatial Catalogue service, and turn it into a virtualised appliance that is hosted in the EGI Federated Cloud. 4. Federate new services: EUDAT Metadata Catalogue and Secure Replication.	
Long term preservation	To be discussed at the H2020 workshop. More details are needed on the use case to suggest a strategy for EGI.	
Services for citizen scientists	4.Federate new services: EUDAT will develop a Simple Store service customised for DRIHM for the citizen scientists use cases. After the service is completed, EGI should consider federating this service into the production infrastructure.	
Data with access restrictions	To be discussed at the H2020 workshop. More details are needed on the use case to suggest a strategy for EGI.	
Data preservation from science gateways	5. Act as a technology provider: By building on long term preservation and open data repositories that may exist in or outside of EGI (see related two use cases), EGI could assemble an API for the developers of science gateways. This API could be used by the gateways to publish data from the gateways in the long term preservation and open access repositories.	
Open Data services	To be discussed at the H2020 workshop. More details are needed on the use case to suggest a strategy for EGI.	
Close compute and data in the cloud	3. New services in the cloud: EGI should bring in 'VA broker' services into the Federated Cloud that could start virtual machines close to the sites where the data used by the virtual machines are located. The broker should access the cloud resources through the OCCI and CDMI interfaces.	

3.3 Cross-cutting questions

The use cases require EGI to develop shared answer to a number of questions that underpin many of the use cases. The EGI Towards Horizon-2020 workshop provides a perfect opportunity to formalise these shared answers. The questions are:

1. EGI resource centres and user communities would benefit from support for the development and implementation of data management policies for hosting data for scientific projects (for both open and restricted access data). What processes, policies and tools should EGI provide to help the setup and implementation of sustainable data management plans?
2. Several use cases require a permanent identifier (PID) infrastructure to make data accessible for the long term. Which PID infrastructure(s) and in what form should EGI support on its production infrastructure?
3. The Unified Middleware Distribution (UMD) of EGI includes the software components that are common to all user communities and are therefore deployed on every site of the production infrastructure. What new software should be included in the UMD to support resource centres addressing cross-cutting needs? (For example data ingestion endpoint, archival storage, data management, data administration, preservation planning, access)
4. EUDAT develops and provides a set of services that serve various scientific use cases. Which EUDAT services and how should be supported in EGI?
5. Many of the emerging scientific use cases require the integration of domain-specific software from external providers into EGI's production infrastructure. How can EGI operate an efficient and scalable software selection and integration process to enable the rapid injection of new software into the production infrastructure?

3.4 EGI strategy

The EGI strategy will be described here in the next version of the paper.