Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 97 (2016) 148 – 152

Procedia Computer Science

CLOUD FORWARD: From Distributed to Complete Computing, CF2016, 18-20 October 2016, Madrid, Spain

Towards European Open Science Commons: The EGI Open Data Platform and The EGI DataHub

Matthew Viljoen[a,*], Łukasz Dutka[b], Bartosz Kryza[b], Yin Chen[a]

[a]*EGI.eu, Science Park 140 1098XT, Amsterdam, The Netherlands*
[b]*Academic Computer Centre CYFRONET, University of Science and Technology AGH, ul. Nawojki 11, 30-950 Kraków, Poland*

Abstract

This paper introduces the EGI Open Data Platform and the EGI DataHub, outlines their functionality and explains how this meets the requirements of EGI end users. The paper also explains how these new services can support the European Open Science Cloud and will fit into the future European Strategy Report on Research Infrastructures (ESFRI).

Peer-review under responsibility of organizing committee of the international conference on cloud forward: From Distributed to Complete Computing

Keywords: Open Science Common; EGI Engage; Data as a Service; Open Data Platform; DataHub; EGI; EGI Foundation

1. Introduction

The recently (April 2016) published ESFRI Roadmap 2016[1] highlights the notion of a *European e-infrastructure Commons* referring to the framework for an easy and cost-effective shared use of distributed electronic resources for research and innovation across Europe and beyond. The concept is outlined by the e-Infrastructure Reflection Group (e-IRG) based on the identification of the need for a more coherent e-infrastructure landscape in Europe. The ultimate vision of the Commons is to reach integration and interoperability in the area of e-infrastructure services, within and between member states, as well as on the European level and globally. This e-infrastructure Commons is

* Corresponding author. Tel.: +31-20-8932-007.
E-mail address: matthew.viljoen@egi.eu

Peer-review under responsibility of organizing committee of the international conference on cloud forward: From Distributed to Complete Computing
doi:10.1016/j.procs.2016.08.294

also a solid basis for building the *European Open Science Cloud* as introduced in the description of the Digital Single Market[2]. The EGI Foundation is responding to this vision by identifying the current blocking issues and analysing how a European Research Cloud could strategically advance its competitiveness by providing research Data as a Service and community-specific tools through a platform that supports the participatory principle of Open Science. As part of its flagship project, EGI-Engage[3], EGI is designing and developing a new Data as a Service (DaaS) offering called the EGI DataHub. The EGI DataHub is based on the Open Data Platform, a distributed data management solution, which provides the backend for efficient data access and sharing on a global scale, with support for open data publishing and access. This paper introduces both the DataHub, the Open Data Platform and their architecture. The authors believe that these solutions will directly contribute to realizing the vision of the ESFRI Roadmap 2016. This will be achieved by providing users with currently lacking but much needed capabilities in federating and integrating computing and data services.

2. EGI and the Open Data Platform

EGI Foundation is a not-for-profit foundation established under Dutch law to coordinate and manage the EGI federation on behalf of its participants – national e-Infrastructures and European Intergovernmental Research Organisations (EIROs). As of April 2016, the EGI federation comprised of 650k CPU cores, 500PB of storage and was serving 46k users.

The EGI-Engage project, funded by the European Commission under the Horizon 2020 programme, was launched in March 2015 with a total budget of 8.7 million Euros for 2.5 years. One of the main objectives of EGI-Engage is to further expand the capabilities of EGI (e.g. cloud and data services) and the spectrum of its user base by engaging with large Research Infrastructures (RIs), the long tail of science (small laboratories and individual researchers), and with industry/SMEs (Small and medium-sized enterprises).

The EGI Open Data Platform, built on OneData[4] technology, is being developed to provide capabilities to publish, use and reuse openly accessible data (including, but not limited to, scientific data sets released into the public domain, publicly funded research papers and project deliverables, and software artifacts and demonstrators coming out of public research projects). Other functionality to be provided by the EGI Open Data Platform will include: policy-based publication, sharing and linking of open data sets; integration of open data access with community portals; data access across federations and support for data provenance. As depicted in Fig.1, the Open Data Platform can be deployed at multiple EGI Federated Cloud[5] sites (or a private computing Cloud site), connecting to various storage systems, including, Lustre[6], Amazon S3[7], Ceph[8], and NFS and other infrastructures. The typical scenario of using One Data Platform can be summarized as follows: A user prepares a data set inside of a Onedata space (a virtual volume). This data set can be of arbitrary size and internal structure in terms of subdirectories and files. The user shares this space with another user who is responsible for ensuring that the data set has appropriate metadata. The sharing functionality is internally provided by Onedata, and requires only exchanging a single token which the owner of the space generates through the user interface. Once the data curator ensures that all relevant metadata is added, the owner creates a snapshot of the entire dataset, which is calculated as a hash of the entire contents of the data set. This ensures that even when the dataset is updated or extended with new files, the handle reference to the dataset points to the exact version which was used when publishing the dataset. Once the data set is published, Open Data Platform OAI-PMH[9] Data Provider service will expose it to OAI-PMH Service Providers such as OpenAIRE[10] on the next scheduled metadata harvest, and the data set will be available and discoverable online.

The platform can be fully integrated with existing EGI Federated Cloud platform allowing users to access data through VMs, running jobs, containers or other application services. Users or research communities can deposit applications and services together with copies of data on to the EGI Federated Cloud. This serves as a pre-staging strategy and is particularly useful for processing large volumes of data, which improves computing performance by avoiding data staging.

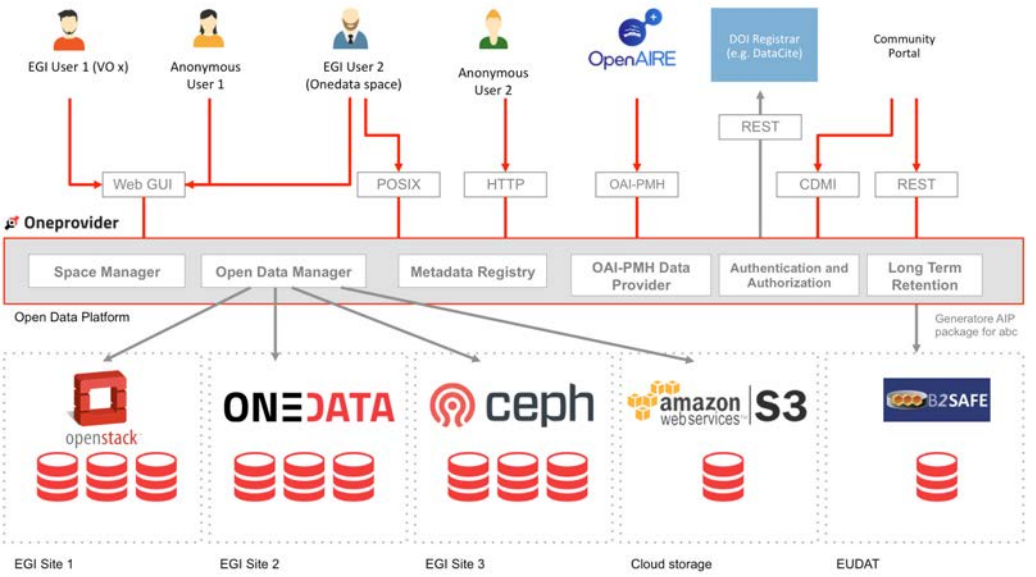

Fig. 1. EGI Open Data Platform architecture.

3. EGI DataHub Service

The EGI DataHub is the end-user service exposing the Open Data Platform functionality, a central point of access for the data collections. The EGI DataHub makes existing large scale open data collections discoverable and available in an easy way for both EGI users and the general public in the case of open data collections.

The DataHub service is based on the Data as a Service (DaaS) delivery model and offers:

1. Access to reference scientific data of public interest. In this context, *reference data* means data that has been created by some communities and there is large interest in using it as an input to visualization or computation. The EGI DataHub allows such data to be discovered and offer long-term access to its replicas to promote ease of use and reuse of the data,
2. Host experimental or temporary scientific data and enable easy access to it by appropriate scientific applications. *Experimental data* means that it may be in the process of being collected, or processed by computations. In case of data collection process, there may be applications monitoring its progress, which could be used to control it. *Data* here could mean datasets, i.e. a collection of data/files/filesets at a level of granularity considered useful to specific user communities.

The DataHub underlying technology, Onedata, has been running in PLGrid Polish Grid Infrastructure since 2014. More recently, it has been used during a collaboration with EGI and the Human Brain Project[11] to demonstrate the fast hosting of data from heterogeneous storage at different service providers. However, the EGI DataHub could be used in the future to augment such dedicated data hosting, for example, if these projects choose to make their data more open and discoverable for other communities.

The EGI Open Data Platform and EGI DataHub are intended to be fully integrated into the existing EGI service portfolio such as the EGI Federated Cloud and EGI Applications Database (AppDB)[12]. The following new functionality is being developed as part of the EGI DataHub – see also Fig. 2:

- Discovery of data via a central portal. This will include a search mechanism plus a rating system which may be based on, for instance, the number of accesses,

- Access to data conforming to required policies which may be: 1) unauthenticated open access; 2) access after user registration or 3) access restricted to members of a user group or Virtual Organization (VO). This access may be via a GUI (e.g. a webpage) or an API (e.g. programmatic access to the data),
- Replication of data from data providers for resiliency and availability purposes. Replication may take place either on-demand or automatically. Replication will require the introduction of a file catalogue to enable tracking of logical and physical copies of data,
- Access to data from the AppDB to enable VOs to associate appropriate data with matching Virtual Appliances,
- Authentication and Authorization Infrastructure (AAI) integration between the EGI DataHub and with other EGI components and with user communities existing security infrastructure

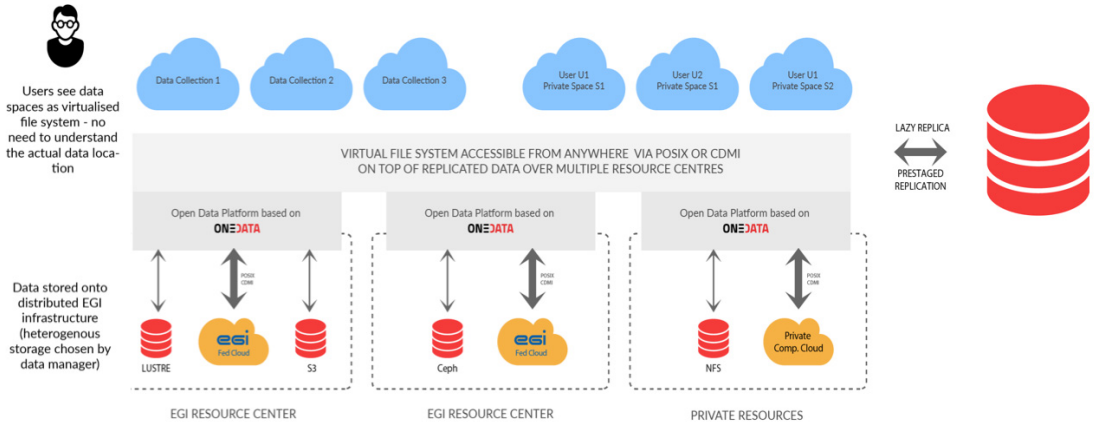

Fig. 2. EGI DataHub Service.

4. Future work and conclusions

This paper presents EGI's e-Infrastructure solutions supporting the European Open Science Cloud. The features of EGI Open Data Platform and the DataHub service are introduced. Currently, these two EGI products are under development, testing and evaluation. A prototype is expected to be released in November 2016 and the final release is expected to be in Autumn 2017. Once the first release version of EGI DataHub and Open Data Platform is published in 2017, the EGI Foundation will continue to ensure its continued operation in production and coordinate future evolution. Sustainability of the national infrastructures hosting data in the federation is a responsibility of the national funding agencies providing such services to the end-user, while EGI Foundation will sustain the operational cost for the DataHub federation services, and innovation is implemented through projects (both national and European). EGI Foundation is also developing a number of pay-per-use models to diversify funding revenues and it is expected that such revenue streams will contribute to the sustainability of services in the future.

Acknowledgements

The development of the Open Data Platform and EGI DataHub service is under EGI-Engage project, which is sponsored by the European Commission's Horizon 2020 Grand scheme (654142).

References

1. ESFRI. Strategy Report On Research Infrastructures: Roadmap 2016. ISBN: 978-0-9574402-4-1, Mar 2016.
2. SWD (2015) 100 final accompanying the document. A Digital Single Market Strategy for Europe COM (2015) 192 final. SWD (2015) 100 final.
3. EGI-Engage: https://www.egi.eu/about/egi-engage/
4. Onedata: https://onedata.org/

5. https://www.egi.eu/infrastructure/cloud/
6. http://lustre.org/
7. https://aws.amazon.com/documentation/s3/
8. http://ceph.com/
9. https://www.openarchives.org/pmh/
10. https://www.openaire.eu/
11. https://wiki.egi.eu/wiki/HBP
12. https://appdb.egi.eu/