

**EGI-Engage**

Open Data Platform in the Production

M4.3

|  |  |
| --- | --- |
| **Date** | 31 August 2017 |
| **Activity** | WP4 |
| **Lead Partner** | Cyfronet |
| **Document Status** | FINAL |
| **Document Link** | <https://documents.egi.eu/document/2839>  |

Abstract

This document presents the status and future plans for moving Open Data Platform to production, including the development and release timelines, summary of experience reports obtained during evaluation within JRA 2.1 activity of EGI-Engage and plans for future applications.

**COPYRIGHT NOTICE**



This work by Parties of the EGI-Engage Consortium is licensed under a Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/). The EGI-Engage project is co-funded by the European Union Horizon 2020 programme under grant number 654142.

**DELIVERY SLIP**

|  |  |  |  |
| --- | --- | --- | --- |
|  | ***Name*** | ***Partner/Activity*** | ***Date*** |
| **From:** | Lukasz Dutka | Cyfronet/WP4 | 23/08/2017 |
| **Moderated by:** | Malgorzata Krakowian | EGI Foundation/WP1 |  |
| **Reviewed by** | Tiziana Ferrari Peter Solagna | EGI Foundation/WP1EGI Foundation/WP5 |  |
| **Approved by:** | AMB and PMB |  |  |

**DOCUMENT LOG**

|  |  |  |  |
| --- | --- | --- | --- |
| ***Issue*** | ***Date*** | ***Comment*** | ***Author/Partner*** |
| **V 0.1** | 28/06/2017 | First draft | Bartosz Kryza / Cyfronet |
| **V 0.2** | 05/07/2017 | Added future plans section. | Matthew Viljoen / EGI |
| **V 0.9** | 07/07/2017 | Final version for review. | Bartosz Kryza / Cyfronet |
| **V 1.0** | 23/08/2017 | Final version | Bartosz Kryza / Cyfronet |

**TERMINOLOGY**

A complete project glossary and acronyms are provided at the following pages:

* <https://wiki.egi.eu/wiki/Glossary>
* <https://wiki.egi.eu/wiki/Acronyms>

**Contents**

[1 Introduction 5](#_Toc491953927)

[2 Open Data Platform production release overview 6](#_Toc491953928)

[2.1 ODP release timeline 6](#_Toc491953929)

[2.2 ODP production release 6](#_Toc491953930)

[3 Preparation for Production and Future Plans 8](#_Toc491953931)

[3.1 Development of Nagios probes 8](#_Toc491953932)

[3.2 Independent review of documentation and security 8](#_Toc491953933)

[3.3 Work in progress and plans 8](#_Toc491953934)

[3.3.1 Business case and value of the EGI DataHub 8](#_Toc491953935)

[3.3.2 Transition to full production readiness 9](#_Toc491953936)

[4 Conclusions 10](#_Toc491953937)

[5 References 11](#_Toc491953938)

**Executive summary**

The EGI Open Data Platform has been developed as part of the EGI-Engage JRA 2.1 activity, by extending Onedata platform with open access features, as identified by a user requirements analysis performed in the first phase of EGI-Engage project and reported in M4.1 report [1].

Open Data platform has been evaluated on various use cases, and the feedback from the use cases has been taken into account when preparing the final release. The results have been described in the deliverable D4.9 “Open Data Platform: Demonstrator, Experience Report and Use Cases” [2].

# Introduction

Nowadays, as large amounts of data are generated, either from experiments, satellite imagery or via simulations, access to this data becomes challenging for users who need to further process them as existing data management solutions often require manual data transfers from the origin to the computational nodes, using command line tools. On the other hand, an important aspect of the research is related to obtaining easy access to open data sets as well as publishing results of research in the form of open data sets. The main problem is that most existing solutions for open access are focused on providing means of storing references to publications, handles (e.g. DOI’s) but lack support for transparent and efficient access to very large datasets (in Petabyte range) by users who do not necessarily have direct access to the storage with the data collections or do not have means to replicate such data sets to near their computing resources.

Open Data Platform provides unified access and management interface to all kinds of users data - small personal files, large research data sets as well as open access collections, all stored on resources provided by EGI resource centers. Users can organize data into virtual folders, called “Spaces”, which can be automatically and transparently distributed and replicated among the storage providers. An important feature of Open Data Platform is to allow users to seamlessly use their data for computation tasks executed on EGI FedCloud infrastructure.

The goal for this milestone was to deploy the Open Data Platform prototype in production environment on resources provided by NGI’s participating in the EGI Federated Cloud. As reported in D4.3, the Open Data Platform demonstrator has been implemented and evaluated on selected use cases and data sets. Currently final preparations are ongoing in deploying a production release version of Open Data Platform based on the feedback from the evaluation, which required some additional development. The production version of the EGI DataHub will be released in the weeks following end of the project, and further sites will be invited to become storage providers for the EGI DataHub.

# Open Data Platform production release overview

## ODP release timeline

Open Data Platform releases have been scheduled based on the main milestones of EGI-Engage project, i.e.

|  |  |  |
| --- | --- | --- |
| M4.1 (M06) | Open Data Platform: requirements and implementation plans  | Preparation requirements analysis and design for Open Data Platform |
| M4.2 (M15) | Launch of call for cross e-Infrastructure case studies  | Launch of first prototype of the platform, including EGI DataHub service |
| M4.3 (M30) | Open Data Platform in the Production  | Release of production ready Open Data Platform service and integration with EGI production services |

## ODP production release

Open Data Platform is based on the Onedata [3] distributed virtual filesystem solution, and extends it with several features necessary for effective management, publishing and usage of large scale open access data sets and collections. ODP’s open access features have been integrated into Onedata system and are available directly from Onedata interfaces. The main features related to open data that have been delivered as part of Open Data Platform include:

* Registration of handle identifiers (e.g. DOI’s)
* Publication of open data sets metadata via OAI-PMH [4] protocol
* Extended metadata support
* Import of existing data
* Extended attributes support
* Detailed administrator manual
* Integrated with EGI Nagios monitoring

The final release has version number 17.06.0 and is available on various Linux operating systems including Ubuntu, Fedora and Centos. It can be downloaded and installed directly from distribution packages [5] or more conveniently using provided Docker images and example setups provided in the form of Docker Compose scripts [6].

Open Data Platform is the basis of an EGI DataHub service, which aims to provision large scale reference data sets to research communities, and is part of EGI Change Management Procedure ensuring that the service upgrades are performed in a controlled and supervised manner.

# Preparation for Production and Future Plans

This section outlines the work that has been carried out to prepare for full production readiness of the EGI Open Data Platform and plans for future work before the end of the project, and afterwards.

## Development of Nagios probes

As with any production service that is relied upon for production usage, monitoring probes need to be developed that are able to track uptime and alert systems administrators to any problems that need to be investigated and fixed.  Such probes are also needed to determine whether SLA targets for uptime are being met.  The Onedata Onezone and Oneprovider components were identified as needing health probes, which were developed in early 2017 [7] and put into production in the EGI monitoring service.

## Independent review of documentation and security

Although the Open Data Platform was primarily designed with open data in mind, data which needs to be protected by VO, group or individual level needs also to be considered. Since the Opendata solution presents a POSIX-like interface with traditional POSIX permissions, it was felt that this needed to be verified separately from Cyfronet, the primary organizing developing Onedata.  Prior to inviting data providers to join an EGI data federation, it was additionally felt to be important for the system administrator setup documentation was verified by installing a new multi-site data federating service, following documentation, and then doing basic security checks of the installation.

The Onedata software maturity has been assessed also by external reviewers in the context of the INDIGO-DataCloud project.  The exercise was considered successful in that basic security checks were carried out and passed and a number of improvements to the setup documentation were identified and carried out.

## Work in progress and plans

### Business case and value of the EGI DataHub

At the time of writing this deliverable, a Service and Design Transition Package (SDTP) is being compiled internally within EGI, as part of the EGI ITSM [8] framework.  This document sets the basis for a transition of the EGI DataHub service into full and sustained production by identifying the current problems those potential target users of a data federation face, alongside the benefits that the new service delivers.  It also sets out the business case including cost estimates of running the data federation, which helps to ensure that the service can be funded sustainably.

Upon completion of the SDTP the EGI DataHub service will transition from Alpha to Beta in the EGI Service Portfolio.  The benefits of joining and using the EGI data federation will be promoted and links to documentation will be made public.

### Transition to full production readiness

Whereas the aim for the delivery of a fully working prototype has been achieved (equivalent to DoD Technology Readiness Level [9], TRL 7), further work is required until the service may be considered fully production ready and suitable for mission critical applications (TRL 9). Most importantly, the early adopters in the project who assisted with testing the Open Data Platform (notably ICOS and Earth Observation) need to complete their testing and deem that the technology is ready for a production launch for their services. This will require a staged rollout, first involving real end users using well defined use cases and then releasing the technology for more general exploitation. At the time of writing this document, Open Data Platform components Onedata is being prepared for inclusion into the EGI Unified Middleware Distribution, thanks to work done within the INDIGO DataCloud project.

# Conclusions

This document presented the current production status of the Open Data Platform developed within JRA 2.1 activity of EGI-Engage, and based on Onedata technology, extending it with open access features.

ODP has been used in several demonstrators including ICOS and Earth Observation, and based on the experience and bugs identified Open Data Platform has been improved. ODP is also a base of the EGI DataHub service, and it will be continuously improved to ensure that users and storage providers associated with EGI DataHub can effectively access selected reference open data sets.

Future plans are focused on deploying and improving the production version of EGI DataHub at <https://datahub.egi.eu> as well as integrating Open Data Platform in further research communities and adding new storage providers with legacy open data sets to the EGI DataHub service.

# References

1. EGI-Engage M4.1 Open Data Platform: Requirements and Implementation Plans, <https://documents.egi.eu/public/ShowDocument?docid=2547>
2. EGI-Engage D4.9 Open Data Platform: Demonstrator, Experience Report and Use Cases. <https://documents.egi.eu/public/ShowDocument?docid=3033>
3. Onedata website, <https://onedata.org>
4. OAI Protocol for Metadata Harvesting, <https://www.openarchives.org/pmh/>
5. Onedata installation, <https://onedata.org/#/home/documentation/doc/getting_started/downloading_onedata.html>
6. Onedata getting started tutorials, <https://github.com/onedata/getting-started>
7. Onedata Nagios plugins, <https://github.com/onedata/getting-started>
8. FITSM website, <http://fitsm.itemo.org/>
9. DoD, Technology Readiness Assessment (TRA) Guidance, <http://www.acq.osd.mil/chieftechnologist/publications/docs/TRA2011.pdf>