EGI-Engage

Open Data Platform: Demonstrator, Experience Report and Use Cases

D4.9

Date	17 August 2017
Activity	WP4
Lead Partner	Cyfronet
Document Status	FINAL
Document Link	https://documents.egi.eu/document/3033

Abstract

This deliverable presents the results from development, provisioning and evaluation of the EGI Open Data Platform prototype, developed as part of EGI-Engage JRA2.1 activity. The EGI Open Data Platform is based on Onedata, an open source distributed virtual filesystem solution, extended with features enabling open data ingress and egress, curation, metadata management and discovery as well as open data publishing.

This material by Parties of the EGI-Engage Consortium is licensed under a Creative Commons Attribution 4.0 International License.

The EGI-Engage project is co-funded by the European Union (EU) Horizon 2020 program under Grant number 654142 http://go.egi.eu/eng

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 Bartosz Kryza	Cyfronet/WP4	28.07.2017
Moderated by:	Malgorzata Krakowian	EGI Foundation/WP1	
Reviewed by	Mario David Yannick Legre Diego Scardaci	LIP EGI Foundation/WP1 EGI Foundation/WP3	
Approved by:	AMB and PMB		17.08.2017

DOCUMENT LOG

Issue	***Date***	***Comment***	***Author/Partner***
V 0.1	28/06/2017	Template and table of contents	Matthew Viljoen / EGI Foundation
V 0.2	30/06/2017	First draft	Bartosz Kryza / Cyfronet
V 0.3	06/07/2017	Added ICOS use case	Matthew Viljoen / EGI Foundation
V 0.9	07/07/2017	Version for review	Bartosz Kryza / Cyfronet
V 1.2	24/07/2017	Addressed feedback from Mario David / LIP, Yannick Legré / EGI Foundation and Diego Scardacci / EGI Foundation	Bartosz Kryza / Cyfronet and Matthew Viljoen / EGI Foundation
V 1.3	28/07/2017	Added new section 4.4 Scalability tests and modified section 5.1 after further input from Mario David / LIP	Bartosz Kryza / Cyfronet and Matthew Viljoen / EGI Foundation

FINAL	2/08/2017	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

Executive summary

The purpose of this document is to describe the main goals and achievements of the EGI Open Data Platform demonstrator activity within JRA 2.1 of EGI-Engage project. The EGI Open Data Platform is a prototype that has been developed based on an initial analysis of requirements from different user communities. This was followed by a comprehensive analysis of existing solutions and a proposal for a prototype service developed over the course of the EGI-Engage project [6].

The EGI Open Data Platform aims at providing a novel technology for federated data access and management of large-scale data sets and collections with direct support for open data use cases such as publishing, discovering and referencing scientific data sets. The EGI Open Data Platform builds on the Onedata distributed virtual file system and adds the features related to open access such as OAI-PMH metadata publishing protocol and DOI registration.

Early testing for a service as ambitious as the EGI Open Data Platform is critical in identifying shortcomings. Within the scope of this activity, demonstrators of the EGI Open Data Platform have been made, and the main issues identified during evaluation have been addressed. The EGI Open Data Platform will be further improved as part of future initiatives. It also forms the basis of the EGI DataHub service, which is the EGI-branded instance of the EGI Open Data Platform.

1 Introduction

Until recently Open Access to research has been mostly considered in the scope of free access to scientific publications such as books, journal papers and conference proceedings. However as more research relies on access to high quality large data sets, including data collected from physical experiments as well as data obtained through pure simulations, it is becoming more apparent the importance of extending the open access also to data sets referenced in the scientific publications.

Currently several efforts are already addressing this challenge, by providing means for data sets indexing and cataloguing, such as DataCite [1] or OpenAIRE [2]. These services rely on established standards such as OAI-PMH [3], which enable them to integrate with the existing platforms for publication metadata harvesting, and identify datasets through globally unique handles such as DOI [4] or PID [5]. However, while these services enable discovery and identification of open data sets, they do not address directly the issue of accessing the underlying data by end users. In case of scientific publications, the typical scenario was to simply download the text document with the publication or view it in the browser, this method is not feasible in case of large data sets required in many fields such as high energy physics or chemistry.

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.

The goal of the EGI Open Data Platform is to fill this gap by providing a high performance, easy to use distributed virtual file systems enabling easy access to open data sets as well as easy provisioning of legacy research data to the community.

2 EGI Open Data Platform Overview

This section presents general concepts and features behind the EGI Open Data Platform, in particular features, which have been added since the previous deliverables M4.1 [6] and D4.7 [7]. The initial overview of requirements from different communities is presented in the EGI Open Data Platform: Requirements and Implementation Plans M4.1 [6].

2.1 Data management

The EGI Open Data Platform is based on the Onedata distributed virtual filesystem solution [8]. Onedata is a globally distributed virtual file system, built around the concept of Space, which can be seen as a virtual folder with an arbitrary directory tree structure. The actual storage space can be distributed among several storage providers around the world, while Onedata ensures transparent access and replication to these resources for users. Each provider gives the users support for each space in a fixed amount and the actual capacity of the space is the sum of all declared provisions. Each space can be accessed and managed through a web user interface (Dropbox-like), REST and CDMI interfaces and command lines as well as mounted directly through efficient POSIX protocol. This gives users several options, the major of which is ability to access large data sets on remote machines (e.g. worker nodes or Docker [9] containers in the Cloud) without pre-staging.

2.2 Metadata management

An important aspect of Onedata is a flexible metadata mechanism allowing for storing metadata in the form of simple key value pairs, as well as entire metadata documents (currently in JSON and RDF formats), which can be attached to data resources and used during indexing and querying. Building on top of this metadata mechanism, Onedata enables users to publish their data as open access content. Onedata supports several protocols and standards for open data such as OAI-PMH for publishing data, Handle system integration for registering DOI or PID identifiers, enabling full open data management life cycle management from ingestion through curation to open access. Furthermore, metadata can be manipulated directly on the POSIX mount of selected spaces using the extended attributes utilities such as `xattr`.

2.3 Open Data publishing

Onedata in addition implements open data publication functionality, by integrating it's concept of shares (data sets which can be exchanged with other users via URL link) with registration of Digital Object Identifiers and metadata dissemination via the OAI-PMH protocol which can be used to directly publish selected data sets as open data and make it discoverable by users of such services as OpenAIRE or DataCite.

2.4 Legacy data import

In use cases where there is a need to provision large legacy datasets, it is possible to configure Oneprovider service to expose such data set directly from the legacy storage without any data migration to another storage[1]. Oneprovider service will run periodically synchronization of files on such storage, and will detect automatically new or updated files and will update its metadata database automatically. This option can be selected when adding new storage to the Oneprovider and has to be performed by Oneprovider administrators.

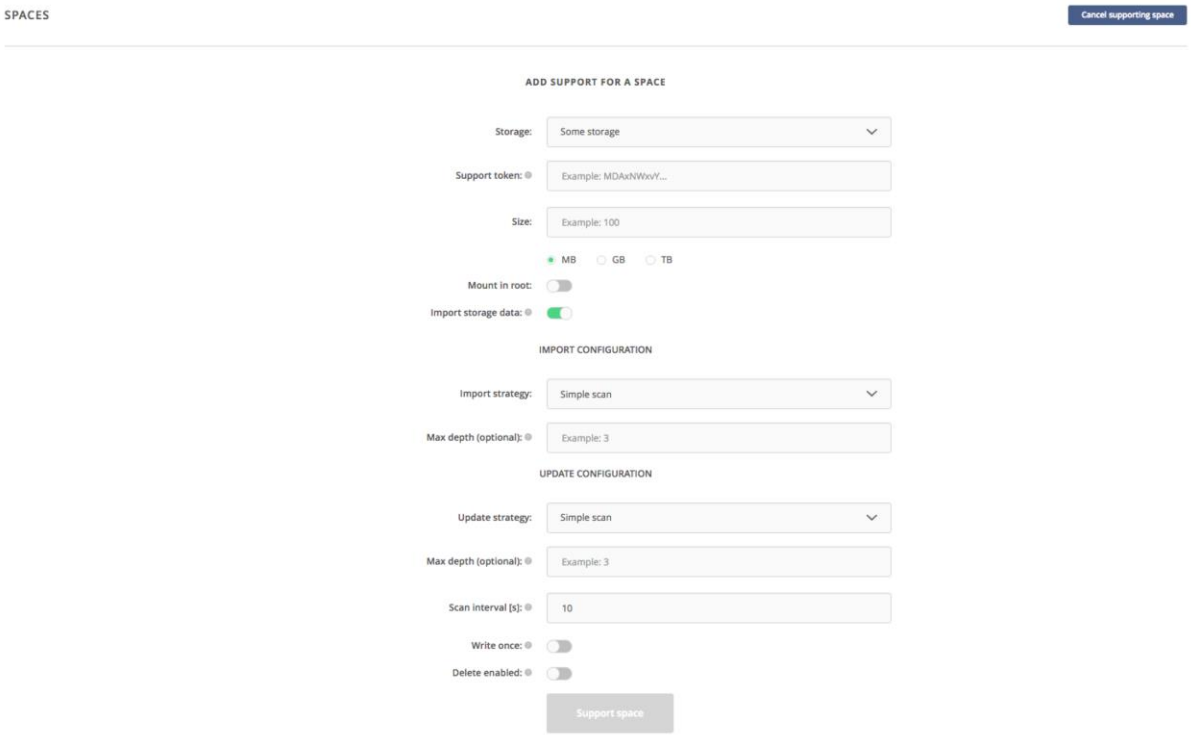


Figure 1 Storage import configuration for legacy data

[1] Available since Onedata version 17.06.0-beta6

Once the storage is configured for the legacy data import, it will be continuously monitored for changes in the data collection (new files, modified files, deleted files) and basic statistics on the scan process will be displayed.

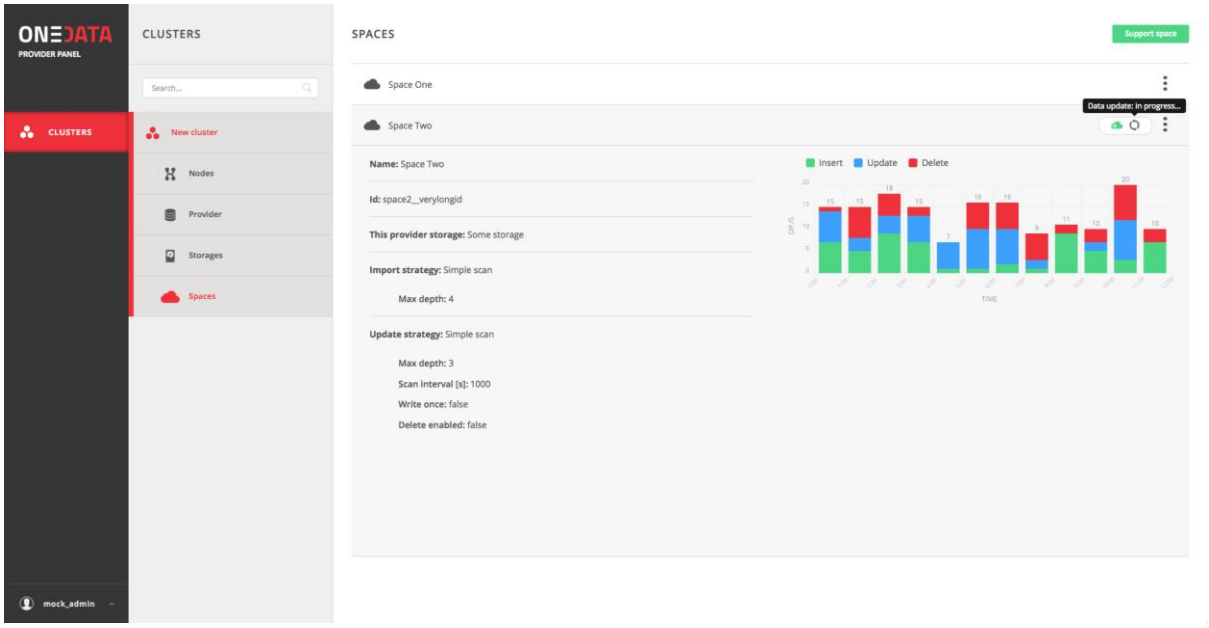


Figure 2 Storage import statistics

3 EGI DataHub

Early on in the project it was realized that it was not enough for EGI-Engage to simply develop a platform for data federation. Many user communities would be unwilling or unable to deploy an independent, isolated data federation. This requires significant system administration knowledge aware of the problems and pitfalls of a data federation. It also means that an independent, domain-specific data federation may not benefit from the open-ness and discoverability that many user communities want, to aid exploitation and re-use of their data. For this reason, it was decided that an important output of EGI-ENGAGE JRA2.1 would be an EGI-branded instance of the EGI Open Data Platform, which is named the "EGI DataHub". This service would be a Data as a Service, designed to fulfil the following roles:

- a service available for multiple communities, regardless of their domain
- a generic central point of discovery for open data
- a means of publishing open data, linked to long-term preservation services such as B2SAFE from EUDAT. This may be done by exposing data from B2SAFE to an Oneprovider by iRODS.
- along with the EGI Fedcloud [10], a well-documented service for bringing data to computing

The EGI DataHub is built on top of the EGI Open Data Platform which, thanks to Onedata technology, enables a wide range of existing storage services to be connected, regardless of their underlying technology (currently supporting natively Amazon S3, Ceph, GlusterFS, OpenStack SWIFT as well as any POSIX compatible storage which can be mounted to the Oneprovider machine). Due to the plugin architecture of the EGI Open Data Platform, integration with the EGI DataHub is seamlessly done by means of installing a Oneprovider server that allows data to be accessed via well-known protocols such as POSIX or web services.

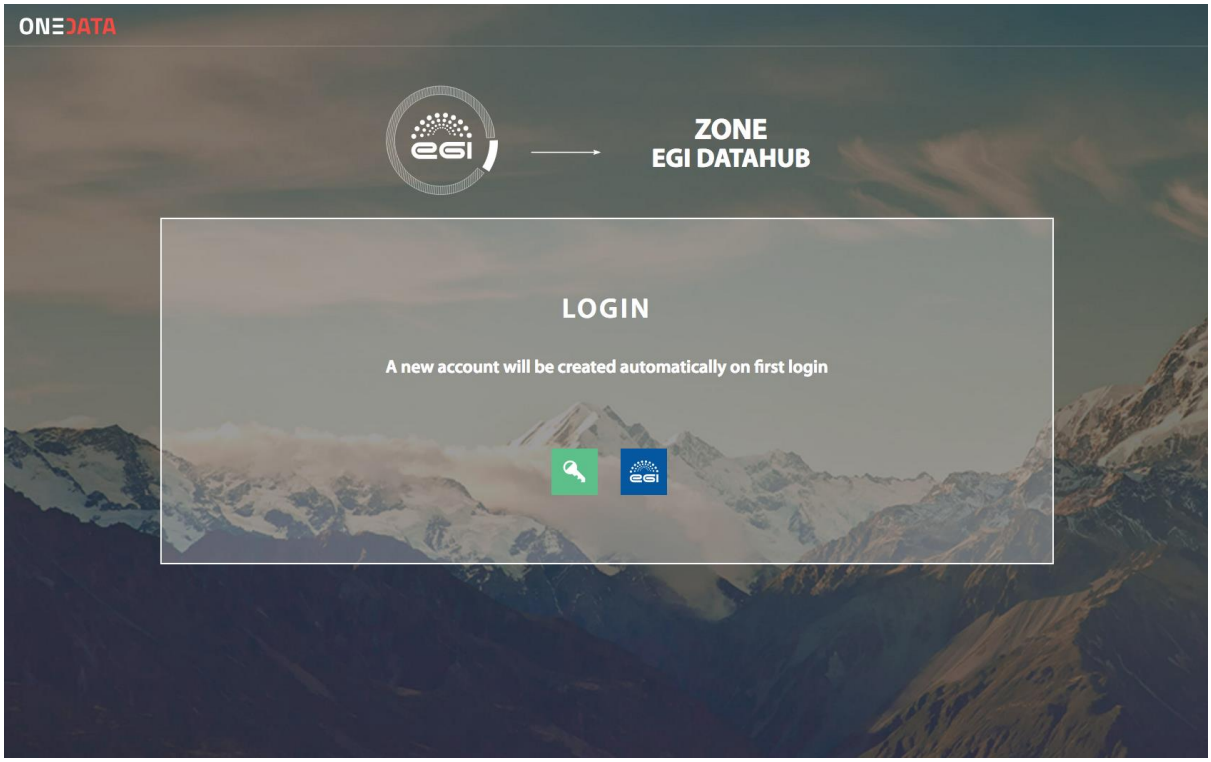


Figure 3 Main DataHub landing page allowing to login using EGI federated IdP (https://datahub.egi.eu)

Currently the EGI DataHub serves a subset of the Sentinel data set. See section 5 for future plans for the EGI DataHub.

4 Demonstrator Use Cases Overview

4.1 Earth Observation PoC

Earth Observation PoC was an initiative by a consortium of organizations including RHEA Group, EOProc, SixSq, Cyfronet, EGI Foundation and AdviceGEO to build a demonstrator enabling easy and efficient access to EO data by end users.

The main objective of this demonstrator was to show how novel Cloud technologies, in particular SlipStream [11] and Nuvla [12] Cloud management systems and Onedata and the EGI Open Data Platform can enable efficient and flexible processing of Earth Observation data products on the example of Sentinel 1 and Sentinel 2 data sets. The goal of the demonstrator was to show how the functionality provided by the proposed architecture can benefit existing Earth Observation data users and not to stress test the selected technologies, as the used data sets were relatively modest.

The specific objectives of this demonstrator included:

- Scalability – data should be processed in parallel by multiple users simultaneously,
- Cloud diversity – data should be located and processed on resources from different Cloud providers,
- Cloud agnostic solution – from the users perspective, the interface should be transparent in terms of the underlying Cloud technology,
- Openness – the image processing services should be treated by users as black-boxes,
- Data location awareness – the data management system should provide detailed information about data location to enable Nuvla to optimize processor deployment,
- Clear benefit for end-users – faster access to data, without the need to download the data first.

The overall architecture of the proposed solution is presented in the figure below. The EGI Open Data Platform was responsible for enabling transparent high-performance access to the Sentinel data sets in a heterogonous Cloud setting, where VMs were managed automatically using Nuvla platform. The core of the platform is constituted by SlipStream acting as Cloud orchestration engine, via which users can schedule and deploy image-processing jobs. Data is stored on resources from different, geographically distributed providers and exposed in a unified way to SlipStream and end user applications via Onedata. SlipStream uses information from Oneprovider instances on data location to schedule jobs in an optimal way, taking into account where the data already is located. User applications when deployed on Virtual Machines automatically have access to the required data sets via Oneclient, which exposes the Onedata virtual filesystem to the user application as a regular POSIX mountpoint. Applications can use the mountpoint also for

writing results, which are then accessible to other application instances for implementing further workflow steps

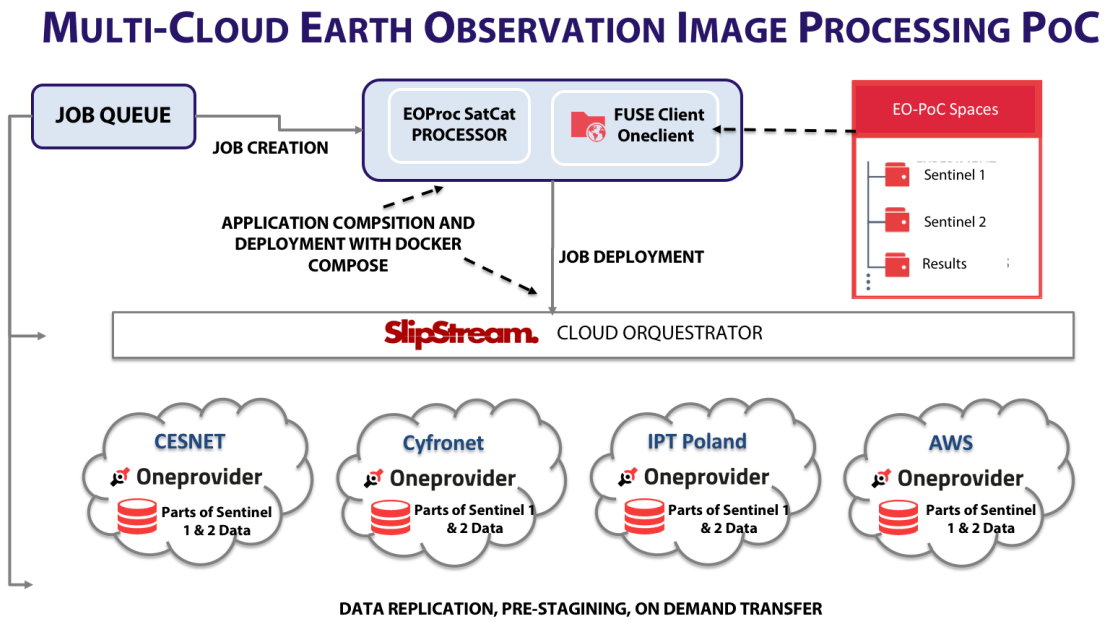


Figure 4 Earth Observation Proof Of Concept prototype by EGI, SixSq and Cyfronet.

The proof of concept included 2 demonstrations:

4.1.1 Sentinel 1 – On-demand processing

This demonstration showed how a user can request evolution of a specific geographical area over time. In particular the demonstration was performed on the evolution of the Panama Canal region using SAR images, which were transformed into animated GIF images showing the landscape change over select period of time. This test involved 26 Sentinel 1 EO scenes, each about 1-4 GB in size.

4.1.2 Sentinel 2 – On-the-fly

The second use case of this demonstrator was focused on processing high-resolution JPEG images from Sentinel 2, which are of inferior quality in comparison to for instance Landsat images. This demonstrator showed that it is feasible to perform on the fly high-quality previews of the Sentinel 2 images.

4.1.3 Evaluation

During the evaluation, Onedata instances were deployed on Amazon Singapore Site, Amazon Frankfurt Site, Cyfronet (Krakow, Poland), CloudFerro (Warsaw, Poland), CESNET (Prague, Czech Republic).

During the evaluation and execution of the demonstrators it was shown that all specific objectives (as stated above) were achieved in both use cases, in particular by the functionality of transparent data access allowing easy processing of data in any location on the Cloud and usage of the detailed data location Onedata API's, allowing Nuvla to select automatically best Cloud resources for processing input data based on their current replication status. Data transfers within Europe observed between sites averaged around 150MB/s, the Singapore-Europe data transfers averaged around 30 MB/s, which was limited by the connection between the Cloud infrastructures.

The test data sets have been also successfully imported into Onedata. The distributed block-based filesystem provided by Onedata proved particularly useful, as most of the files were large (>1GB) ZIP archives. Using Oneclient, it was possible to list the contents of each ZIP files and it's metadata without the necessity to download the entire ZIP archives, but only their headers, and the meteorological application was able to process only interesting files, without having to download not needed files. This functionality has been completely transparent to user applications. Furthermore the REST API enabled easy integration of higher-level application logic with the Onedata data management functionality, in particular data replication between data centres.

To conclude, this evaluation was considered to be successful by the Earth Observation community involved in this work, both from the point of view of the functionality demonstrated and by the transfer speed achieved and sample file sizes used.

4.2 ICOS

4.2.1 Introduction

The Integrated Carbon Observation System (ICOS) [13] is a pan-European research infrastructure for quantifying and understanding the greenhouse gas balance of the European continent. ICOS has been working closely with the EGI Engage project within the context of developing the 'Footprint Tool for Atmospheric Stations'. This is a web-based tool designed to help users explore the sensitivity to European emissions at specific ICOS atmospheric sites. This tool uses input from meteorological analysis as well as emissions data to determine the greenhouse gas footprint over time. This is done by making use of the Stochastic Time-Inverted Lagrangian Transport (STILT) model [14].

The ICOS use case has been implemented using services from both EGI and EUDAT. Data is stored on EUDAT B2SAFE [15] and users interact with the ICOS Carbon Portal which instantiates a number of VMs in the EGI Federated Cloud to host the a number of virtual machines fulfilling different functions.

Over the course of the project, two distinct workflows have been considered, the first employing the EGI Open Data Platform to provide input data via a federated storage while the output data is written to and NFS server, and the second employing the EGI Open Data Platform for both input and output data from the workflow.

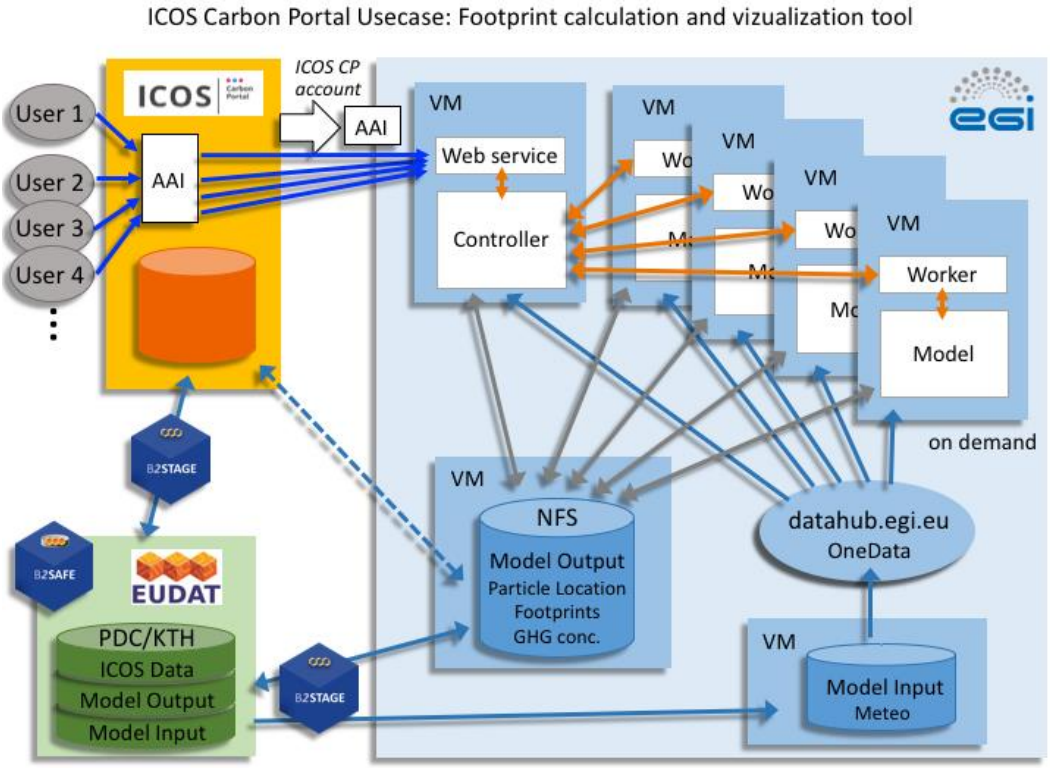


Figure 5 ICOS Footprint workflow #1

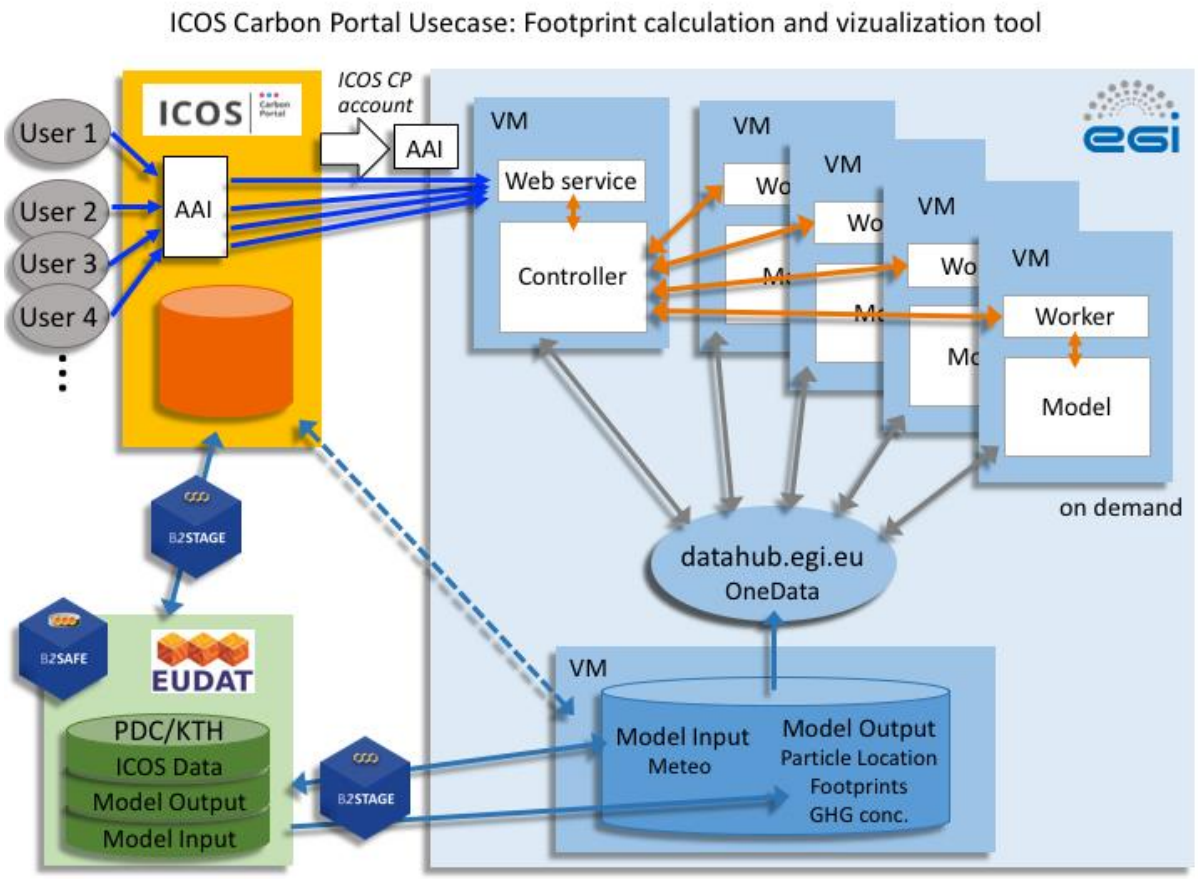


Figure 6 ICOS footprint tool workflow #2

The workflow components running on EGI FedCloud virtual machines (VMs) are as follows.

1. A VM hosting the web interface for requests of model runs and visualization of the results. This VM also runs the controller for load balancing and VM orchestration. The controller distributes the jobs to the required number of VMs and passes input parameters to one or more worker node VMs (see below)

2. One or more VMs hosting the worker node for running the model and sending back-log files. This VM hosts the model itself, and due to the parallelisable nature of the workflow, may have multiple instances running concurrently. An Oneclient runs on each of these VMs exposing the data to the model.

3. A VM running Oneprovider serving as the model input (meteorological aspects and emissions maps). For workflow #2 this is also used for the model output. Access to this Oneprovider is mediated by a Onezone. During the testing this Onezone was hosted as part of the usecase but, in the future, the plan is for this Onezone to be the EGI-hosted EGI DataHub.

4. (only workflow #1) a VM running an NFS server serving as the model output which is later transferred to EUDAT B2SAFE by B2STAGE.

This use case has considered NFS and Onedata as the means for distributing data to and from the VMs. The advantages of Onedata as identified by ICOS are:

- Onedata enables federation not only across VMs on a cloud hosting site, but also across different cloud hosting sites.
- Onedata allows for the data to be discovered and accessed by anyone, using a Onezone such as the EGI DataHub.

4.2.2 Use case implementation of the Open data Platform

All three EGI Open Data Platform Onedata components were installed (Onezone, Oneprovider, Oneclient) according to the online documentation. It was found that the solution worked for large input files, but was problematic when writing small output files. This was found to be due to Onedata not efficiently supporting instant creation of large numbers of small files (tens of thousands) at the time of the latest release available at the time (Onedata 3.0.0-rc11). For this reason, workflow #2 was suspended until the small file problem was addressed (in the later Onedata version 17.06.0-beta2[2]), and work continued with workflow #1.

Another important issue identified by the ICOS use case testing was a problem related to incompatibilities between different Onedata components (Oneclient and Oneprovider were running rc14 but the Onezone was running 3.0.0-rc11). This feedback has now been addressed, as it is recognized that it is impractical to have all components in a data federation on the same version at all times, and to schedule upgrades at the same time.

At the time of writing this deliverable, the EGI DataHub is being updated to the latest version, which addresses both of these issues identified by the ICOS testing, and once done, testing of the use case will resume using workflow #2. If successful, further testing will continue by using the EGI Open Data Platform to provide data for both input and output steps of the workflow.

4.3 Performance tests

In addition to the tests of functional aspects, which were covered by the community use cases tests, performance and stress tests have been performed, partially using input from the community use cases but also as part of other activities and initiatives. Due to the relevance of these results to the use cases, they are included as part of this deliverable.

The main goal of these tests was to determine Onedata performance depending on the storage technology used as a backend for user spaces, which directly impacts users experience when

[2] As of version 3.0.0-rc16 Onedata versioning scheme has changed, current versions start with a year and month followed by minor release number, e.g. 17.06.0

access their data over Onedata POSIX protocol. The results below show performance for the following storage types with Oneclient working in Direct IO mode (transferring data blocks directly from and to the storage):

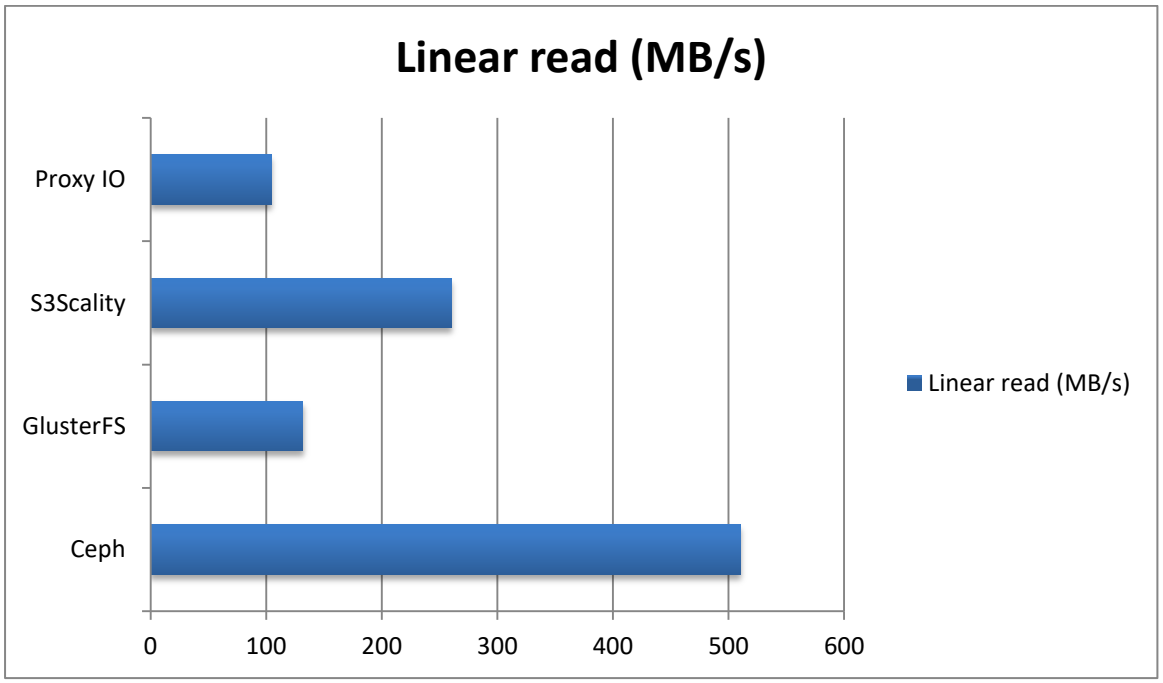
- Ceph
- GlusterFS
- S3Scality

Also for comparison test has been performed showing the performance when accessing data in Proxy IO mode (Oneclient has no direct connection to the storage and all data block transfers have to go via Oneprovider).

All tests have been performed on 10Gbit network with Oneprovider instances deployed on 8vCPU, 16GB virtual machines with SSD local storage for Oneprovider database.

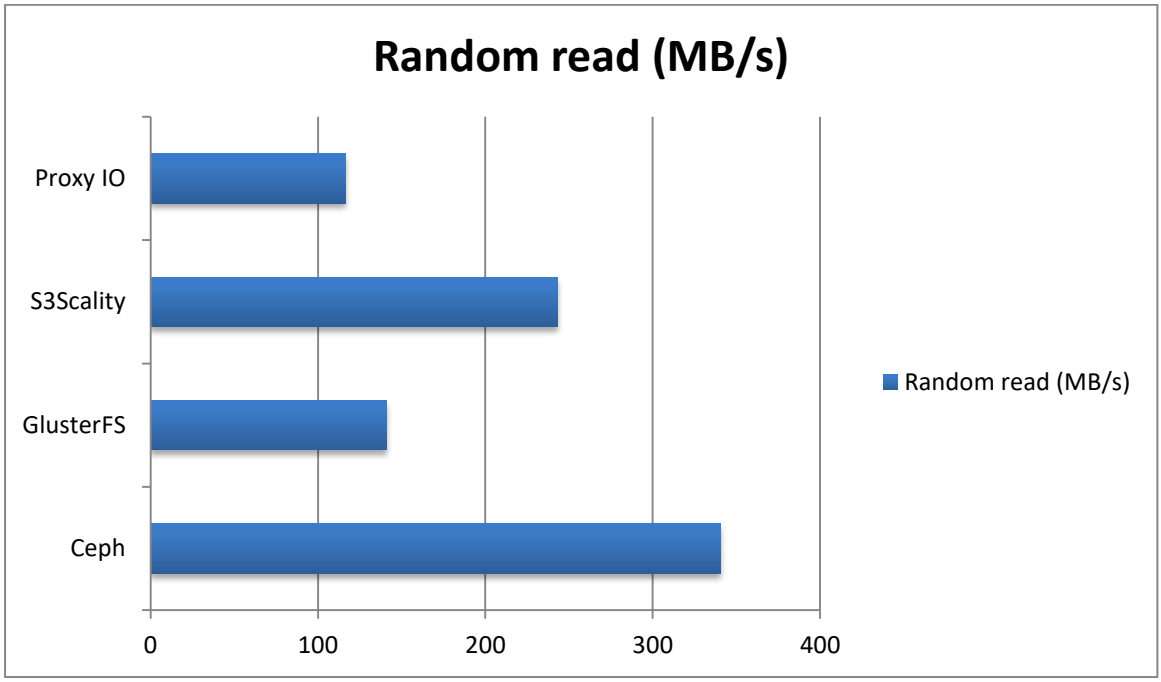
4.3.1 Linear read

This test was performed by running 8 parallel processes over one Oneclient mountpoint reading 50GB data set in 50MB consecutive blocks.



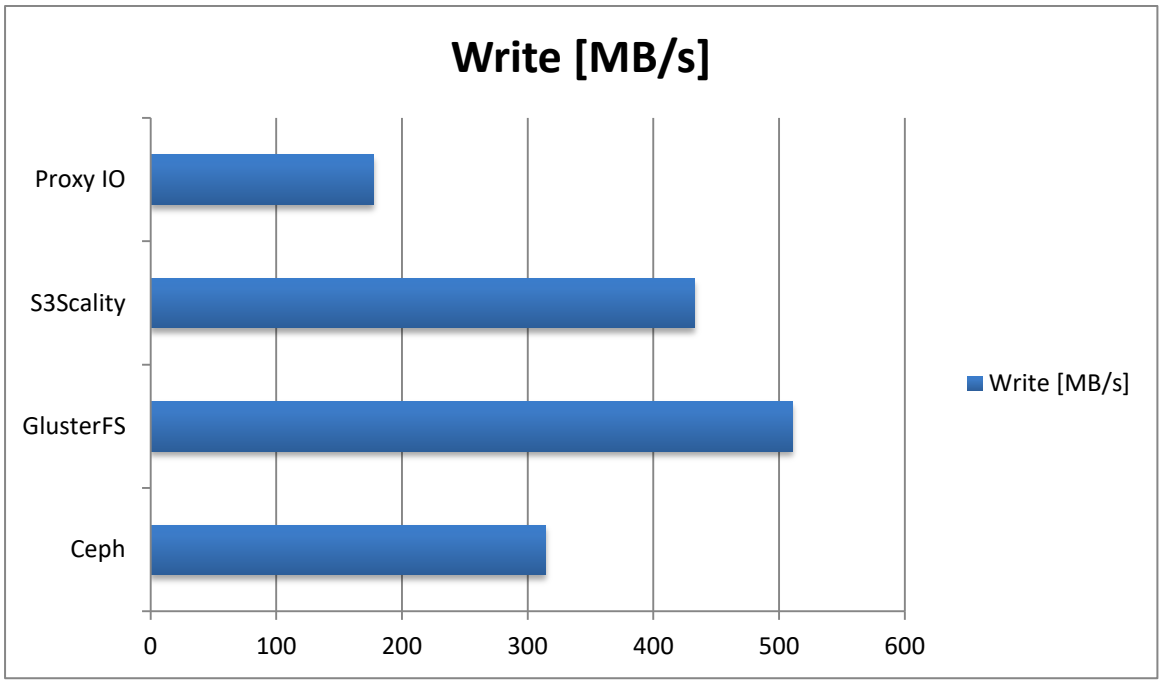
4.3.2 Random Read

This test was performed by running 8 parallel processes over one Oneclient mountpoint reading 50GB data set in 50MB blocks in random order.



4.3.3 Write

This test was performed by running 8 parallel processes over one Oneclient mountpoint writing 50GB data set in 50MB blocks order.



The results show that the data access performance via Onedata virtual filesystem can be still very efficient while maintaining complete transparency over the actual storage backend used to support user spaces. Selection of actual storage technology usually involves, in addition to performance, several other aspects such as already existing legacy storage solution, the number of machines necessary to run storage service, the cost of the service. Typically object based storages such as S3 are cheaper than high performance storage solutions, which have to be deployed on additional Virtual Machines.

4.4 Scalability tests

In addition to performance tests, at the time of writing this deliverable, Onedata is also being tested for handling of large data volumes, in order to ensure that large data collections can be managed by the EGI Open Data Platform. One important use case is the import of the CloudFerro mirror (http://www.cloudferro.com/en/eostats/) of Earth Observation Sentinel-1 data set, which consists of 625TB in around 750,000 files. This test makes use of the new Onedata feature, which enables automatic scanning of legacy storage and importing information about existing data (files, directories and their metadata) and exposing such storage to the end users. The results of these tests are planned to be available by the end of August 2017, and will be published as an update to the EGI-Engage milestone report M4.3 or as a separate report.

5 New use cases and potential future users of the EGI Data Federation

Two use cases have been at the forefront of testing the EGI Open Data Platform - ICOS and Earth Observation Proof of Concept. However, EGI has been in discussion with others communities who also see the potential benefits of a federating data solution to their work and are interested in Onedata as a solution.

5.1 INDIGO DataCloud

The INDIGO DataCloud project [18] is a software development project that has been funding development of Onedata, at the same time as development of Onedata funded by EGI-Engage. Onedata is an INDIGO component in addition to being the enabling technology of the EGI Open Data Platform. As a result, there have been mutual benefits to the two projects from the complementary work in development of different aspects of the technology.

A number of community partners of INDIGO DataCloud (e.g. Molecular Dynamics Simulation, EMSO, CMS, CTA) have provided their requirements and tested Onedata against their requirements, recognising the benefits of the solution. These communities have been involved in testing on the INDIGO Pilot Preview Testbed, which will cease to exist at the end of the INDIGO DataCloud project. While it is possible that some of these communities will continue to test Onedata as part of a follow-on project [17], others may chose to go into full production with Onedata. A number of these communities have expressed an interest in joining an EGI data federation or a federation run by others (e.g. INFN) or indeed a federation run by themselves. An interesting possibility here (if there is sufficient demand for such functionality) is to enable a Oneprovider to be registered in multiple Onezones. This would enable a data provider to be able to move seamlessly between data federations, or to be a member of multiple federations at the same time to allow for maximum exploitation of their data by different communities and different services in multiple infrastructures

5.2 EOSCPilot

The Findable, Accessible, Interoperable and Re-usable (FAIR) principles of data are at the centre of the European Open Science Cloud and easy access and exploitation of open data is at the centre of this. Three of the initial science demonstrators within the EOSCPilot project (PanCancer, DPHEP and the Photon-neutron science community) are interested in the functionality of Onedata and are interested in an externally managed data federation such as the EGI DataHub. EGI will work with these communities and assist them with testing. If the demonstrators are a success, it is likely that they will wish to make use of the data federation that EGI is building.

6 Plan for Exploitation and Dissemination

Name of the result	EGI Open Data Platform
DEFINITION	
Category of result	6.1.1.1.1 Software & service innovation
Description of the result	The EGI Open Data Platform aims at providing a novel solution for open data management, giving the researchers similar experience and ease of use as with commercial data management and file synchronization solutions, while providing means for seamless publication and access to open data from any location, either from personal laptop or virtual machine running in the cloud.
EXPLOITATION	
Target group(s)	6.1.1.1.2 The main target groups of the EGI Open Data Platform are: • RIs • international research collaborations • long-tail of science • industry/SMEs
Needs	6.1.1.1.3 The main community requirements fulfilled by the EGI Open Data Platform have been identified within M4.1 milestone deliverable and include: • Publication of open research data based on policies • Make large data sets available without transferring them completely • Enabling complex metadata queries • Integration of the open data access data management with community portals • Data identification, linking and citation • Enabling sharing of data between researchers under certain conditions • Sharing and accessing data across federations • Long term data preservation • Data provenance

How the target groups will use the result?	6.1.1.1.4 The EGI Open Data Platform will be used by target communities to manage, process, share and disseminate open data, which are input or output essential to their research activities.
Benefits	6.1.1.1.5 The main benefits for the user communities include: • Unified high-performance data access and management system • Easy access to remote data sets from Virtual Machines and containers via standard POSIX protocol • Direct support for open data publishing • Secure data sharing between users and across federations • Easy exposing existing large data collections from legacy storage
How will you protect the results?	6.1.1.1.6 The EGI Open Data Platform as well as its underlying technology, Onedata, is fully open-source components licenses under Apache 2.0 license.
Actions for exploitation	Currently the EGI Open Data Platform is being integrated into the EGI operational services via procedure PROC19 [16]. Once this is complete, selected data centres federated in EGI infrastructure will deploy Onedata and register it with EGI DataHub service to provide a distributed open data environment for researchers.
URL to project result	https://onedata.org https://datahub.egi.eu https://github.com/onedata/onedata
Success criteria	6.1.1.1.7 The main success criteria for this product are: • Number of storage providers provisioning storage space to the EGI DataHub service • Number of open data sets published via the EGI Open Data Platform platform • Number of communities use cases integrated with the EGI Open Data Platform • Number of open data sets accessed by external users
DISSEMINATION	
Key messages	• With the EGI Open Data Platform, users can access, store, process and publish data using global data storage backed by computing centres and

	storage providers worldwide, • The EGI Open Data Platform focuses on instant, transparent, POSIX-compliant access to distributed data sets, without unnecessary staging and migration, allowing access to the data directly from your local computer or worker node, • The EGI Open Data Platform makes the process of open data publishing effortless by supporting Handle based identifiers (e.g. DOI) and OAI-PMH protocols
Channels	6.1.1.1.8 The EGI Open Data Platform will be disseminated through several channels including: • EGI website and newsletter • Scientific publications • Open science conferences
Actions for dissemination	• Live demonstration during DI4R conference in Krakow, September 2016 • Hands-on session during DI4R conference in Krakow, September 2016 • Hands-on session during 3rd ENVRI+ Week in Prague, November 2016 • Presentation during Workshop on Cloud Services for Synchronization and Sharing (CS3), January 2017 • EGI Community Forum, May 2017 • Presentation during INTERNATIONAL CONFERENCE ON COMPUTATIONAL SCIENCE, ICCS , June2017 in Zurich
Cost	6.1.1.1.9 No additional costs except for the already allocated within EGI-Engage will be necessary.
Evaluation	6.1.1.1.10 The dissemination results will be evaluated using several means: • Number of unique visits to the EGI DataHub website • Number of new users registering on the EGI DataHub service • Number of citations of publications related to EGI Open Data Platform

7 Conclusions and Future Plans

This document presents the final EGI Open Data Platform version and two representative demonstrators, namely ICOS and Earth Observation Proof of Concept, that have been prepared based on Open Data Platform. Issues identified during the evaluation have been addressed and integrated into the final release.

From the technical requirements and development perspective, all required features of Open Data Platform (as identified in [6]) have been delivered as part of EGI-Engage JRA 2.1, using development effort either from EGI-Engage or from INDIGO DataCloud projects. This includes federated data management with transparent access over POSIX protocol (INDIGO), advanced metadata management (ENGAGE), support for DOI registration (ENGAGE) and publishing of open data set metadata via OAI-PMH protocol (ENGAGE) and support for automatic importing of legacy data directly via Onedata storage synchronization mechanism (ENGAGE).

Although the EGI-Engage project ends in August 2017, EGI is very likely to continue to work with both communities covered in this deliverable. The EGI DataHub has a well-defined business case, providing a central service where open data may be discovered and making use of Onedata for scalable and easy access and processing of data by cloud services. Future development of Onedata is made possible by existing funding channels through EOSCpilot project and through accepted projects: EINFRA-21-2017 [17] project eXtreme DataCloud (XDC) and EINFRA-12-2017 [19] project EOSC-hub. As a result, we anticipate that when the EGI DataHub achieves full production status, it will be fully usable and available for exploitation by communities such as the ones outlined in this deliverable.

Appendix I. REFERENCES

1. DataCite - https://www.datacite.org/
2. OpenAIRE - https://www.openaire.eu/
3. OAI Protocol for Metadata Harvesting - https://www.openarchives.org/pmh/
4. Digital Object Identifier - https://www.doi.org/
5. PID Consortium - http://www.pidconsortium.eu/
6. EGI-Engage M4.1 Open Data Platform: Requirements and Implementation Plans - https://documents.egi.eu/document/2547
7. EGI-Engage D4.7 Open Data Platform First Prototype - https://documents.egi.eu/document/2976
8. Onedata project website - https://onedata.org
9. Docker website - https://www.docker.com/
10. EGI FedCloud - https://www.egi.eu/services/cloud-compute/
11. SlipStream platform website - http://sixsq.com/products/slipstream/
12. NuvLa website - https://nuv.la
13. ICOS RI project website - https://www.icos-ri.eu/
14. Stochastic Time-Inverted Lagrangian Transport model information page - http://www.stilt-model.org/
15. B2Safe service website - https://www.eudat.eu/b2safe
16. EGI PROC19 description - https://wiki.egi.eu/wiki/PROC19
17. EINFRA-21-2017 Platform-driven e-infrastructure innovation - http://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/topics/einfra-21-2017.html
18. INDIGO DataCloud - https://www.indigo-datacloud.eu/
19. EINFRA-12-2017 Data and Distributed Computing e-infrastructures for Open Science, http://ec.europa.eu/research/participants/portal/desktop/en/opportunities/h2020/topics/einfra-12-2017.html