EGI-Engage

Open Data Platform: Requirements and Implementation Plans

M4.1

Date	28 September 2015
Activity	WP4
Lead Partner	CYFRONET
Document Status	FINAL
Document Link	https://documents.egi.eu/document/2547

Abstract

This document provides an overview of requirements from selected communities interested in publishing, using and reusing Open Data. The communities' requirements have been collected using custom questionnaires and the summary of these findings has been described in this document, along with an overview and comparison of data management technologies related to open data provision. Technological gaps were identified by comparing requirements to available technologies and a recommendation for technology selection and development priorities is proposed.

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

DELIVERY SLIP

	Name	*Partner/Activity*	*Date*
From:	Bartosz Kryza	CYFRONET	6 Aug 2015
Moderated by:	Sandro Fiore	UNISALENTO	10 Aug 2015
Reviewed by:	I external review: Sandro Fiore Giacinto Donvito II external review: T. Ferrari, Y. Chen, M. Viljoen, B. Jones, J. Shiers	UNISALENTO INFN EGI.eu, CERN	10 Aug 2015 Sep 2015
Approved by:	AMB and PMB		09 Oct 2015

DOCUMENT LOG

Issue	*Date*	*Comment*	*Author/Partner*
V.0	13 Jul 2015	First draft on structure	Y. Chen, L. Dutka, B. Kryza, T. Ferrari
V.1	29 Jul 2015	First draft	B. Kryza, Y. Chen, L. Dutka
V.2	6 Aug 2015	First version for external review	B. Kryza, L. Dutka, Y. Chen
V.3	15 Sep 2015	Second version for external review	B. Kryza, L. Dutka, Y. Chen
V.4, 5	Oct 2015	Third version for external review	B. Kryza

TERMINOLOGY

A complete project glossary is provided at the following page: http://www.egi.eu/about/glossary/

Acronym	*Definition*
CDMI	Cloud Data Management Interface
CTA	Cherenkov Telescope Array
DOI	Digital Object Identifier
EGI	European Grid Initiative
EML	Ecological Modelling Language
GSI	Grid Security Infrastructure
HBP	Human Brain Project
ICT	Information & Communication Technologies
NGI	National Grid Initiative
OAI-PMH	Open Archives Initiative - Protocol for Metadata Harvesting
ODP	Open Distributed Processing
OWL	Web Ontology Language
PDB	Protein Data Bank
POSIX	Portable Operating System Interface
RDF	Resource Description Framework

REST	Representational State Transfer
SKOS	Simple Knowledge Organization System
URL	Uniform Resource Locator
VOMS	Virtual Organization Membership Service

Contents

1 Executive summary

This milestone report presents the processes and results of the investigation on communities' requirements for the new EGI Open Data Platform.

In order to prepare the report, a special requirement questionnaire template has been prepared, focusing on open data access aspects of the community's domain specific data management issues. The report has been sent to representatives of communities and based on their feedback a summary of requirements highlighting major data management characteristics and specific open data access issues have been gathered.

After collection of the requirement reports, they have been summarized in a tabular form in order to compare various aspects of data management between the communities. The comparison revealed high heterogeneity in data management patterns and policies between communities. Furthermore, there is a high diversity in typical data object sizes between communities (ranging from KB to TB), different data and metadata formats and data access and transfer technologies (from using simple tools like rsync to custom data management tools developed within communities).

In order to proceed with implementation of the Open Access Data platform prototype, a set of common requirements was identified and compared with analysis of state of the art technologies available currently. Based on the analysis, selection of technology for the basis of EGI-Engage has been proposed and a list of gaps, which need to be developed, has been identified. These requirements include conditional release of data into the public, making large datasets available for processing without migrating them, support for complex metadata queries, integration of Open Access Data platform with community portals, support for unique data identification (e.g. through DOI) and referencing, enabling selective data sharing between researchers and research groups, long time data preservation and in some cases data provenance. Prototype activities will be based on the further development of the onedata solution given the existing capability of providing in one solution CDMI standard support, data sharing between individual and groups and the capability of federating data hosted by distributed heterogeneous data centre. However, several new features have to be developed in order to support the identified user requirements.

Interoperability with other data infrastructures – primarily EUDAT – will be ensured by the adoption of standard interfaces and protocols for data discovery, access and transfer. Specifically, interoperability with EUDAT will be carried on in task WP4.3 under the guidance of use cases from four selected Research Infrastructures: ELIXIR, ICOS, BBMRI and EISCAT-3D.

2 Introduction

2.1 Purpose

The purpose of this document is to identify major requirements of the research communities with respect to open data access, which would enable and foster publication of research data and results in an open manner, under certain restrictions depending on some domain specific policies. The document's goal is to select and propose technology or set of technologies, which will serve as the basis for the EGI's Open Data Access platform. Based on these analysis, recommendations for a plan for implementing such a platform is made.

2.2 Open Data Access platform vision and goals

A new challenge for existing computing infrastructures in Europe is how to better support fostering dissemination and exploitation of open distributed research data through implementation of data management plans allowing data preservation, distribution and provision over extended periods of time to any researcher and scientist. The vision of Open Data Access platform for EGI, realized through EGI-Engage, is to define an architecture and prototype a technical platform to address such needs.

Eventually, Open Data Access platform will extend the EGI Federated Cloud with distributed data management and access capabilities, including data replication and bringing the computation to where data is located.

Furthermore, through the SME engagement activities, the open data solution will support the mission of the European Big Data Value: "… *foster commercial and social added value based on the intelligent use, management and reuse of data sources in Europe, through a combination of Research and Innovation, legislative and deployment actions. This will lead to world class applications, new business opportunities involving SMEs, increased efficiency of the private and public sectors and to an (open) data friendly policy and business environment*"[1].

2.3 Our Problems

To accomplish the tasks, we have encountered many difficulties, including but not limited to:

- How to better understanding of research community requirements in the requirement collection process.
- How to efficiently collect information with limited resources.
- How to provide useful analysis of information gathered, while sometimes information is missing and difficult to be obtained.

[1] http://cordis.europa.eu/fp7/ict/language-technologies/data-value-chain-home_en.html

- How to define valuable recommendations, which will drive the development stage.

2.4 Scope of the investigation

The collection of the requirements and subsequent analysis performed in this report had the following major focus points:

- Focus on the Open Data Platform technology, which will be designed to foster the discovery, dissemination and exploitation of open data, also addressing the problem of co-location of data and computing for big data processing. The Open Data Platform will provide a distributed data management solution allowing communities to manage data according to their Data Management Plans, including publishing data to selected communities or publicly within certain time frames (e.g. after 1 year from creation). It was initially planned to base Open Data Platform on the onedata data management solution[2], due to its support for federated data management, and this document is investigating if the communities' requirements are matching technological possibilities of the onedata platform and other technologies;
- Focus on data, computation, and use of e-Infrastructures;
- Focus on the requirements of EGI user communities, in particular those contributing to the EGI-Engage Competence Centers;
- Focus on technology applicable to the EGI Federated Cloud.

2.5 Structure of the report

The rest of the report is arranged as follows. Section 3 presents the methodology used in the requirements collection process. Section 4 introduces the communities and their use cases. Section 5 reports the analysis of the requirements and findings. Section 6 gives an overview of the state-of-the-art technology for Open Data. Section 7 identifies the gaps between requirements and technologies, and gives the recommendation on the priorities for developments. Finally, Section 8 concludes this work.

[2] onedata data management solution: http://www.onedata.org

3 Methodology

3.1 Design and Use of Templates in the Requirement Collection Process

Using Templates to Collect Requirements for the Open Data Platform

Following the guidance given by the generic template (refer to the EGI wiki site[3]), we have planned the following activities in order to gather requirements for the Open Data Platform： 1) scope the investigation 2) prepare the template, 3) collect the information 4) review and get approvals from the communities.

Scope the investigation. We first met with the technology development team to understand the general capabilities that the Open Data Platform should deliver.

Prepare the template: The current requirements, challenges and expectations of the communities interested in making their data public within the EGI framework were gathered. Based on the technical expectations, the questions to enquire communities were:

- What kind of data is managed? What formats and sizes characterise the data managed by the community?
- What are the life cycles of data created within the community?
- What are the current data management technical platforms used within the community?
- What is the preferred way for users outside of the community to access public community data?
- What are the potential use cases for public users to access community data (e.g. verification, simulation, visualization, etc.)?

Identify target communities we focussed on the EGI user communities, in particularly, those being technically supported in EGI-Engage. We held a workshop in the EGI User Community Forum 2015 and invited users to discuss the issues[4].

Collect requirements, review and approval: information was initially gathered from available material like design documents and presentations, and then validated by the concerned user communities.

Our Experiences and Known Problems

[3] https://wiki.egi.eu/wiki/Requirement_Collection

[4] Towards an Open Data Cloud session in EGI User Forum 2015, Lisbon, 19-22 May. Description and slides: https://indico.egi.eu/indico/sessionDisplay.py?sessionId=80&tab=contribs&confId=2452

The information collected using a customised template is included in appendix.

Here, we share some of our experiences in the requirement collection process:

- Missing information from the user communities. In our case, since open data platform is a relevant new technology, many investigated user communities haven't addressed the related issues in their practices. As a result, there is missing requirements information from some of the communities. We have gathered information from communities that have provided information and analysed their common needs. The analysis can be re-evaluated when more information becomes available in future.
- Some of the answers were delivered too late to be considered for the requirement collection activity due to various reasons, however they will be considered in the future system evolution as much as possible.
- Using the template to extract desired information from available materials (and later to obtain approval from the community) is an efficient approach to get a quick (and approximated) result when the resource for requirement collection is limited. Using this approach, we were able to efficiently accomplish the task within relevant small time scale and staff budget. It is also useful when external community doesn't have resources to support requirement collection, e.g., to attend interviews or to fill complicated questionnaires. However, there are limitations in applying this approach, e.g., the information is extracted from different materials written by different people in the community. Since different people in the community have different views or knowledge of community activities, it would be hard to obtain approvals from (a single member of) the community.
- Requirements evolve over time. However, the collection of requirements is valid. It depicts a snapshot of a continuing evolving blueprint of a future system. The system design based on the requirements collected shall be extensible and be able to address future needs.

4 Research Communities

This section presents an overview of the research communities considered in the requirements collection, grouped into 4 categories: biological and medical sciences, environmental and Earth sciences, agricultural sciences, astronomy and astrophysics. Technical requirements are presented in detailed in complementary documents that are available at:

https://documents.egi.eu/document/2546.

4.1 Biological and Medical Sciences

4.1.1 Human Brain Project

The goal of the Human Brain Project (HBP) is to accelerate our understanding of the human brain by integrating global neuroscience knowledge and data into supercomputer-based models and simulations. This will be achieved, in part, by engaging the European and global research communities using six collaborative ICT platforms: Neuroinformatics, Brain Simulation, High Performance Computing, Medical Informatics, High Performance Computing, Neuromorphic Computing and Neurorobotics.

For the HBP Neuroinformatics Platform, a key capability is to deliver multi-level brain atlases that enable the analysis and integration of many different types of data into common semantic and spatial coordinate frameworks. Because the data to be integrated is large and widely distributed, an infrastructure that enables “in place” visualization and analysis with data services co-located with data is requisite. Providing a standard set of services for such large data sets will enhance data sharing and collaboration in neuroscience initiatives around the world.

In brain research, open data currently is not a black and white issue – it cannot simply open or close all research datasets. It needs granularity to be explicit about what is open, when and for what purpose and the community needs time and convincing to understand the need for sharing data. Many brain researchers are willing to share data, however, there are obstacles e.g., data are expensive to produce (intellectual capital may include experimental design, acquisition cost, and time); rewarding currencies are intellectual advances, publications and citations, but there are no clear rewards or motivation for providing data completely free of any constraint.

4.1.2 Structural biology (MoBRAIN)

MoBRAIN lowers barriers in accessing modern e-Science solutions from micro to macro scales. MoBRAIN builds on high throughput computing and cloud compute services and on the expertise provided by the WeNMR community[5], N4U[6] and technology providers (NGIs and other institutions,

[5] http://www.wenmr.eu

[6] https://neugrid4you.eu

OSG). This initiative aims to serve its user communities, related ESFRI projects (e.g. INSTRUCT) and, in the long term, the Human Brain Project.

By integrating molecular structural biology and medical imaging services and data, MoBRAIN will kick-start the development of a larger, integrated, global science virtual research environment for life and brain scientists worldwide. The mini-projects defined in MoBRAIN are geared toward facilitating this overall objective, each with specific objectives to reinforce existing services, develop new solutions and pave the path to global competence centre and virtual research environment for translational research from molecular to brain.

4.1.3 BBMRI

Thousands of biobanks in Europe have been collecting data, samples and images of millions of individuals in different stages of their lives, during disease and after recovery. Biobanking is currently evolving from local repositories to a pan-European RI, the BBMRI-ERIC. The BBMRI Competence Centre facilitates the implementation of big data storage in combination with data analysis and data federation by integrating technologies from community projects, EGI and other e-Infrastructures.

Typical services provided by biobanks for their users include sequencing, genotyping and expression profiling. However not all biobanks provide even these services, while several users would benefit from using their own applications and algorithms on data stored in biobanks. This not only requires that data is provided in the form of Open Data Access platform, it also requires that security concerns are met (data is not open to the public by default) and metadata about how the data was sequenced is made available.

From the data exchange perspective, BBMRI-ERIC is committed to FAIR principles[7] (Findable, Accessible, Interoperable, Reusable), with accessibility the limited by privacy protection of patients and donors given the nature of data in BBMRI-ERIC infrastructure. This implies that access is only provided to the authorized users, i.e., typically researchers who work on research projects that have been reviewed by a competent ethical review board.

4.2 Environmental and Earth Sciences

4.2.1 EMSO

EMSO is a large-scale **European Research Infrastructure** in the field of environmental sciences. EMSO is based on a European-scale distributed research infrastructure of **seafloor observatories**[8] with the basic scientific objective of long-term monitoring, mainly in real-time, of environmental processes related to the interaction between the geosphere, biosphere, and hydrosphere, including natural hazards. It is composed of several deep-seafloor observatories, which will be deployed on specific sites around European waters, reaching from the Arctic to the Black Sea

[7] Data FAIRport,http://datafairport.org/

[8] http://www.emso-eu.org/infrastructure/what-are-ocean-observatories.html

passing through the Mediterranean Sea, thus forming a widely distributed pan-European infrastructure.[9]

Ocean observatories are sources of interdisciplinary ocean data across time and space. Ocean observatories provide power and communication connections for sensors to allow a sustained interactive presence in the Ocean. Sensor systems can either be attached to a cable, which provides power and enables data transfer, or they can operate as independent, stand-alone benthic and moored sensor platforms. Data, in both cases, can be transmitted in real time either through fibre-optic cables or through cable and acoustic networks that are connected to satellite-linked buoys, back to shore data centres and the Internet. *(from EMSO Brochure)*.

EMSO data is working towards completely open up its datasets. EMSO uses a data publisher for earth and environmental services, PANGAEA[10], which harvests and archives metadata from various domain scientific data providers and allows users to search and publish their own metadata.

4.2.2 LifeWatch

LifeWatch is a part of the European Strategy Forum on Research Infrastructure (ESFRI) and virtual laboratory facilities for biodiversity research. Many countries and institutions collaborate and contribute results to the LifeWatch community. Within the EGI-Engage a LifeWatch Competence Center is being established, to provide support, training and use case analysis.

The LifeWatch infrastructure is designed around the collection of data from sensor networks that is gathered, transferred and processing through cloud compute services. The typical data flow in LifeWatch (on the example of VLIZ data flow) includes: data gathering by vessels, storing data on the server, backup generation in the cloud, data analysis, storage of the data generated from the analysis, users can access and analyse data through a web portal. Data in LifeWatch are openly accessible to biodiversity researchers.

4.3 Agriculture

4.3.1 Agrodat.hu

The AgroDat.hu project aims to create an agricultural information system using big data technologies. High-volume data about crops and environmental conditions is collected constantly by field sensors. This data is then analysed to discover hidden relations and to suggest appropriate actions. An interactive portal is used to share information with producers and to provide an integrated search tool in agricultural databases.

4.3.2 agINFRA

The agINFRA[11] project, supported by the Agriculture Information Management Standards of the Food and Agriculture Organization of the United Nations (AIMS FAO)[12] and the CIARD[13] global

[9] From http://www.emso-eu.org
[10] http://www.pangaea.de/
[11] http://aginfra.eu

initiative, introduces a set of recommendations for the agri-food research community about data management, sharing and dissemination. Additionally, these recommendations aim to provide a framework for the research community of European agri-food research institutions that need to follow the H2020 Open Access[14] mandate and share their metadata with their thematic aggregator in order to publish them in OpenAire[15] [16].

The produced data is usually stored on personal computers, carrier media and/or restricted access servers, and only shared locally or within collaborative research teams, mostly by sending files directly to collaborators or using restricted file sharing platforms. While open access document repositories only promote self-archiving of already published works, research data collections are not easily shared and a culture of open research data needs to be promoted.

4.4 Astronomy & Astrophysics (A&A)

4.4.1 CTA

The Cherenkov Telescope Array (CTA) is a large array of Cherenkov telescopes of different sizes and deployed at an unprecedented scale. It will allow significant extension of our current knowledge in high-energy astrophysics. The aims of the CTA can be broadly grouped into three main themes, the key science drivers:

1. understanding the origin of cosmic rays and their role in the Universe,
2. understanding the nature and variety of particle acceleration around black holes,
3. searching for the ultimate nature of matter and physics beyond the Standard Model.

CTA use cases involve different users at different stages. Guest Observer, who needs to get access to proprietary data through the VO Space web portal. A pipeline user needs to access raw data to perform data reduction tasks; the produced data needs to be archived. Depending on the data policy, after 1 year all proprietary data will become public, so it will be available through the data access service to the public. After this automatic publication process, publishing of higher-level products to the Virtual Observatory servers is made possible.

4.4.2 LoFAR

LoFAR will be the first large radio telescope system wherein a huge amount of small sensors are used to achieve its sensitivity instead of a small number of big dishes. For the astronomy application, LoFAR is an aperture synthesis array composed of phased array stations. The antennas in each station form a phased array, producing one or many station beams on the sky. Multi-beaming is a

[12] http://aims.fao.org
[13] http://www.ciard.info
[14] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf
[15] https://www.openaire.eu
[16] From www.aginfra.eu

major advantage of the phased array concept. It is not only used to increase observational efficiency, but is important for calibration purposes. The phased array stations are combined into an aperture synthesis array. The Remote Stations are distributed over a large area with a maximum baseline of 100 km within the Netherlands and 1500 km within Europe.

LoFAR data become public accessible since March 2015. Before that there have been limited usage of the data (only by a few scientists). They expect data access can increase after the open data action.

4.4.3 CANFAR

Advanced Network for Astronomical Research (CANFAR) is a computing infrastructure for astronomers. CANFAR aims to provide to its users easy access to very large resources for both storage and processing, using a cloud based framework. CANFAR allows astronomers to run processing jobs on a set of computing clusters, and to store data in distributed data centers.[17] The main objectives of the community supported by CANFAR include:

- Manage large astronomical and astrophysical data sets,
- Allow users to share the data sets between European and Canadian infrastructures,
- Provide means for data set querying using FITS metadata,
- Enable running computations on large data sets.

With respect to open data, two main use cases have been identified. First, all user data is made public after one year from generation, after which data should be replicated between EGI and CANFAR infrastructures. Second, users who want to access public CANFAR data should be able to do so through the web portals using FITS metadata contained in observation files and indexed in external databases.

[17] From http://www.canfar.net/about

5 Analysis and Findings

The goal of this section is to give a summary of the selected communities' requirements and illustrated in https://documents.egi.eu/document/2546, related to open data access as well as to identify a set of their common requirements.

5.1 Summary of the Communities Requirements

This section presents summarized information from requirement documents obtained from community representatives. The summaries are presented in a tabular form, focusing on key aspects of communities requirements related to open data access issues.

Cells are shared when no information is available to date and requirements are still under discussion. As several research collaborations embrace multiple research groups, requirements and policies may change for each of these.

5.1.1 Open access policies

Community name	Data access policy	Is usage tracking required?[18]	How long the data should be preserved?	Other usage restrictions
Human Brain Project	Tier 0 - Unrestricted: All metadata and/or data freely available (includes contributor, specimen details, methods/ protocols, data type, access URL)		Long-term[19]	Tier 1 - Restricted use: Data developing analysis algorithms Tier 2 - Restricted Use: Data available for non-conflicting research questions[20] Tier 3 - Restricted use: Full data available for collaborative investigation, joint research questions
MoBRAIN	End results are deposited into public database like PDB, EMDB or cryoEM	Data are owned typically by the end users. They become public if/when	Long-term (10 years)	Most data is private to investigators, but interpreted data are usually made public in

[18] This question was identified in a later stage of requirement collection, we keep it in the report for future reference purposes.

[19] The actual period of data preservation is community specific, but it is necessary to enable communities to sign an SLA declaring for what time period (e.g. 10 years) the data will be maintained along with relevant metadata.

[20] Such data has limited availability only to users whose research is not competitive to each other

Community name	Data access policy	Is usage tracking required?[18]	How long the data should be preserved?	Other usage restrictions
		deposited in public databases (Presumably, usage/download must be tracked). For data, which is made public, it would be nice to have the tracking stats.		open databases upon publication. It is planned that simulation data could be opened.
BBMRI	Access is only provided to the authorized people, i.e., typically researchers who work on research projects that have been reviewed by a competent ethical review board.	Traceability is required, which requires identifiers with time-stamping support for both data and samples, and ability to identify subsets generated by queries on data in specific time.	Long-term	Oviedo Convention (ETS 164), the Helsinki Declaration, the OECD Guidelines for Human Biobanks and Genetic Research Databases (HBGRD) (OECD, 2009) or the Directive 95/46/EC on the Protection of Personal Data.
EMSO	Working towards opening all datasets, in Europe, uses the INSPIRE and GEOSS data sharing principles.	This has not been raised as a critical requirement yet. But citation and corresponding tracking is more interesting for the community	Long-term	Currently certain datasets are subject to embargos of different lengths, or are simply unavailable for technical reasons
LifeWatch	Data is open access		Long-term	For managers IPA system at IFCA (similar to LDAP). Users TBD.
Agrodat.hu	Open access		Long-term	
agINFRA	Data is open access		Long-term	
CTA	Will be open accessible when operational (first time in this field)		Long-term (over 30 years)	
LoFAR	Data that has passed the proprietary period becomes public and can be retrieved by anyone		Long-term	None
CANFAR	Typically public after 1 year		Long-term	

5.1.2 Data characteristics

Community name	Typical object sizes	Overall collection size estimate	Data formats	Current data management technologies used	Open data access protocols
Human Brain Project	Each image will typically range from 1-10TB	O(10PB) currently, it will grow to O(1Exabyte) within next 5-10 years	Brain scans are stored in a form of: series of bitmaps, VTK[21] (for 3d rendering), HDF5[22], TIFF/JPEG at origin, convert to HDF5 From the data structure point of view a single scan is either file or a directory of files.	BBIC	HTTP queries
MoBRAIN	Raw NMR data (1-50MB per sample) Processed NMR 100MB) Analysed NMR data (several GBs) Cryo-EM data (up to terabytes)	There are no overall collections	PDB [23](Protein Data Bank), Text files (only for the 3D structures). Raw data are stored typically in proprietary formats.	MoBrain (WeNMR site) are not managing data repositories. As best they have storage elements on the grid.	Usually journals require the processed data (read structures) to be deposited in public databases like the PDB or cryo-EM database (EMDB).
BBMRI	Sample sizes vary	34,000,000~46,000,000 samples	Unstructured, federated data base with sematic data support (triple store systems and translation of ontologies)	BBMRI-ERIC Directory, Sample broker, Sample locator, Ontology translation service, Sensitive data processing and sharing platform	
EMSO	Between several MBs to several GB per data set, depending on the	There are no overall collection	NetCDF and ODV (Ocean Data View). Use of SWE standards is	PANGAEA/MOIST OAI-PMH are use for metadata harvesting and integration;	FTP

[21] http://www.vtk.org/wp-content/uploads/2015/04/file-formats.pdf
[22] https://www.hdfgroup.org/HDF5/
[23] http://www.rcsb.org/pdb/home/home.do

Community name	Typical object sizes	Overall collection size estimate	Data formats	Current data management technologies used	Open data access protocols
	instrumentation and configuration of the observatory.		being encouraged.	Ifremer/PANGAEA SOS are used for real-time data acquisition; MOIST Opensearch for data searching	
LifeWatch	Zooscan (~4GB/sample), VPR (150kb/image, 10GB/h), Flow cytometer (~ 200MB/sample), Acoustic fish telemetry (~ 25MB/month), Multibeam echosounder (sediment 10Gb), Water column (100Gb/day), Sediment profiler imaging (1Gb/image), Acoustic bat recorder (1MB/sec, 0.5Gb/night)	Zooscan (432 GB/year), Flow cytometer (1TB/year), Sediment profiler imaging (130Gb/year), Bird tracking with GPS (several GB/year)	Text based files (CSV[24], XYZ[25], others) Images/Videos stored in JPG[26]	Rsync is used for synchronizing data. GPFS as system to store the data. NFS to export file system to web servers. Bacula: software for preservation. Planning to test OneData.	HTTP
Agrodat.hu		High-volume	Images	A new big data server farm with hierarchical storage with noSQL database, GPGPU cluster for processing the raw data, Hadoop servers, etc. Open Stack with Ironic.	GSM network and M2M communication enabled SIM cards
agINFRA	~10KB	~1PB	XML[27], MCPD[28] (Germplasm data)	Custom solution	HTTP

[24] http://www.ietf.org/rfc/rfc4180.txt

[25] http://openbabel.org/wiki/XYZ_%28format%29

[26] https://en.wikipedia.org/wiki/JPEG

Community name	Typical object sizes	Overall collection size estimate	Data formats	Current data management technologies used	Open data access protocols
CTA		>1000PB (target size)	FITS[29], RAW, ROOT[30], JSON[31], XML, BSON[32]	CTA Computing Grid (CTACG) for simulation runs, DB & Archive System for storage	CTA Scientific Gateway
LoFAR	1 datacube (~TB) LOFAR telescope allows up to 488 subbands, (GBs) Observational data 60 Gbps (650 TB/day)	>19PB (3PB grows each year)	datacubes (3D data)	LOFAR standardized pipelines	web data portal
CANFAR	~1TB/one night observation	~1PB	FITS	VOSpace	HTTP, FTP

5.1.3 Metadata characteristics

Community name	Metadata format	Metadata storage (files, databases)
Human Brain Project	Some metadata are included in the file but most of them are stored in JSON and XML files.	Files
MoBRAIN	No standards, mainly simple text files	Files, but the PDB must be storing data in databases.
BBMRI	ICD-9[33], ICD-10, SNOMED CT[34], UMLS[35]	
EMSO	ISO 19115, DIF or NetCDF, and an extended version of Dublin-Core.	Files, in PANGAEA archive

[27] http://www.w3.org/XML/

[28] http://genbank.vurv.cz/ewdb/asp/multicrp.htm

[29] https://en.wikipedia.org/wiki/FITS

[30] https://root.cern.ch

[31] http://www.json.org

[32] http://bsonspec.org

[33] https://en.wikipedia.org/wiki/International_Statistical_Classification_of_Diseases_and_Related_Health_Problems#ICD-9

[34] http://www.ihtsdo.org/snomed-ct

[35] http://www.nlm.nih.gov/research/umls/

Community name	Metadata format	Metadata storage (files, databases)
LifeWatch	Ecological Metadata Language[36] (EML)	Various, generated and stored at external data providers' sites
Agrodat.hu	INSPIRE SensorML	Cassandra (noSQL DB)
agINFRA	RDF[37], OWL[38], XML, SKOS[39], OAI-PMH[40]	Files, RDF Triple stores
CTA	Astronomical FITS standards and VO standards	Standard parsing/ingestion in the DB and move to Archive
LoFAR	HDF5	Files
CANFAR	FITS	

As indicated in the tables above, the 10 research communities considered in this document are characterized by high heterogeneity of data management patterns, usage of various data formats and metadata standards. Sometimes at present these are lacking data management policies (e.g. when and what data should be made public). This requirements document will be periodically updated to include additional open data management information as it becomes available.

5.2 Identification of the Common Requirements

This section presents an attempt of extrapolating from the detailed requirements questionnaires received from communities into a small set of key requirements for the open access data management platform, which will be developed within EGI-Engage.

5.2.1 REQ1: Publication of open research data based on policies

Many communities require that some of the data obtained from experiments or simulations should be made available to the public based on various conditions. For instance in case of agricultural data (agINFRA), most data is public immediately. For astronomical data (CTA, LoFAR, CANFAR) data is private to the Principal Investigator for 1 year, after which the data should be made publicly available.

Other communities may require even more complex open access policies, such as HBP, where we need granularity to be explicit about what is open, when and for what purpose, then gradually develop the culture of loosening these restrictions.

[36] https://en.wikipedia.org/wiki/Ecological_Metadata_Language
[37] http://www.w3.org/RDF/
[38] http://www.w3.org/2001/sw/wiki/OWL
[39] http://www.w3.org/2004/02/skos/
[40] https://www.openarchives.org/pmh/

5.2.2 REQ2: Make large data sets available without transferring them completely

For several communities (such as HBP, CANFAR, LoFAR), which produce very large data sets in large files (>100GB) it is not convenient to migrate data to other sites in order to make them public. This can include transferring selected subsets of data sets or directly mounting external datasets using virtual file systems. The latter could be important for legacy applications, requiring POSIX style access to data. Thus a method for directly accessing the data from the source sites has to be provided.

5.2.3 REQ3: Enabling complex metadata queries

Due to the nature of the data generated and processed by the considered communities, an essential aspect of the data management system for open access data is to support specific metadata used within the communities. The main problem is that metadata standards are very heterogeneous across communities. For instance astronomical communities use the FITS standard where metadata on each data set are stored in the file header (which consist of multiple key/value pairs), which are further indexed in relational databases. Other communities, such as agINFRA or HBP plan to use complex ontologies based on RDF or OWL standards, requiring specification of semantic queries in languages such as SPARQL.

5.2.4 REQ4: Integration of the open data access data management with community portals

Many of the analysed communities give access to their resources, including data, through custom portals prepared according to domain specific requirements, and whose users are accustomed to in terms of user interface, terminology and data querying features. This includes VOSpace portal for astronomical communities or HADDOCK portal for communities involved in biomolecular research. It would be important to integrate open access data management software directly with the portals, so that public users can use the same domain specific interface to search for public as well as restricted data sets, depending on their access rights.

5.2.5 REQ5: Data identification, linking and citation

Most communities require that open access data is provided with information on how to uniquely identify and cite the data used for further research. In particular, data owners should be able to generate persistent citable links to data. For many use cases it would suffice to use DOI identifiers, however some may require more complex solutions (e.g. LifeWatch plans to adopt an enhanced Life Science Identifier).

Furthermore, in some cases, data is not available in data repositories, but can be generated on demand by certain services (e.g. HBP). In such case a link should convey information about how to generate the data.

5.2.6 REQ6: Enabling sharing of data between researchers under certain conditions

For communities where data is not automatically public since its inception, in some cases it could be beneficial for data owners (such as Principal Investigators in case of astronomical communities) to share certain datasets with researchers whom they trust and would like to collaborate with, without requiring them to register to the data owners' infrastructure. This sharing could be then controlled

by the open data platform with certain restrictions, e.g. for how long certain data set is available, and to which users.

5.2.7 REQ7: Sharing and accessing data across federations

In many cases, the communities leverage several infrastructures resources and store their data in multiple infrastructures simultaneously. For instance astronomical communities use the VOSpace infrastructure, however for certain purpose, such as access to EGI's computational resources the need exists to easily and securely access data between the infrastructures.

5.2.8 REQ8: Long term data preservation

Several communities, including agINFRA, LOFAR, MoBRAIN and HBP require long-term preservation of data. This entails ensuring that infrastructures storing their data have long term data preservation policies in place. Furthermore, not only raw data has to be preserved, but also metadata related to this data, otherwise most data becomes useless once metadata is lost, or the connection between metadata and data (i.e. links or identifiers) becomes lost.

5.2.9 REQ9: Data provenance

In some communities (e.g. HBP), an important issue is that of data reproducibility, i.e. information on how to regenerate data sets or when data is not stored at all, but only produced by certain services on demand. This requires the data management platform to store somewhere information, for instance at the metadata level, on workflows and input data necessary to generate certain datasets. These are unfortunately very specific to each community and their data and metadata standards.

6 Review of State-of-the-Art technology for Open Data

This section provides an overview of existing technologies with potential to support open data use cases of EGI communities. The main focus of this section is on technologies and tools, which enable efficient sharing, transfer and remote access to large data sets (with open or managed access) either obtained directly from experiments or generated through simulations.

6.1 ownCloud

ownCloud[41] is an open-source framework for creating self-managed file hosting services (Figure 1), similar to Dropbox, i.e. sync-and-share. It enables to maintain full control over data location and transfers, while hiding the underlying storage infrastructure, which can be composed of multiple storage resources.

The main features of ownCloud include abstracting file storage available through directory structures or WebDAV, file synchronization between various operating systems, built-in calendar/task/address book functionality, user group administration, sharing of files using public URLs, online text editing, viewers for various file formats, support for external Cloud storage services (e.g. Dropbox or Google Drive).

[41] https://owncloud.com/whitepapers/

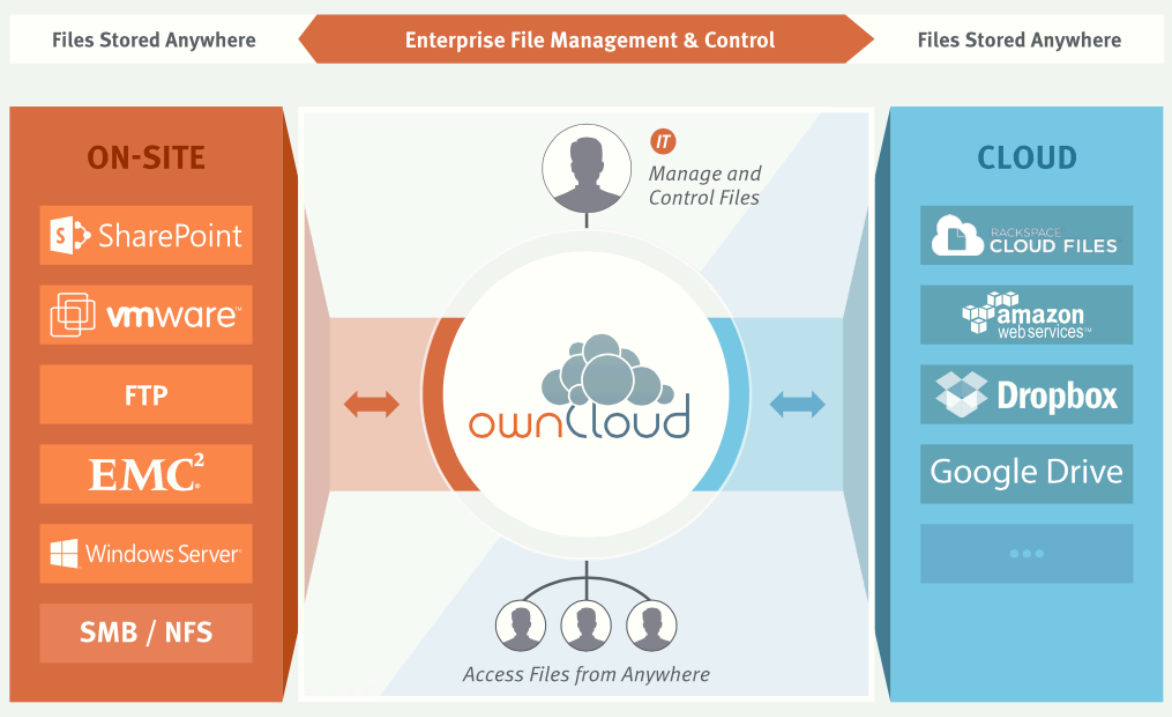

Figure 1 ownCloud overall functionality

From the point of view of open data, ownCloud supports publication of links to data sets (files) using public URLs. However, ownCloud is more focused on consumer applications, i.e. support for HPC in terms of optimized file transfers or remote read/write POSIX access are not available.

6.2 iRODS

The Integrated Rule-Oriented Data System (iRODS)[42] is an open source data management software used to manage and take control on users' data regardless of the device used to store data (Figure 2). It's main features include data discovery using a triple based metadata catalog, support for data workflows, with a rule engine allowing any action to be initiated by any trigger on any server or client in the grid, secure collaboration and data virtualization, allowing access to distributed storage assets under a unified namespace, and freeing organizations from getting locked in to single-vendor storage solutions.

Metadata in iRODS may be attached to files, users, groups, collections (iRODS equivalent of sub-directories), and resources (data containers [e.g., a hard drive]). Each iRODS zone contains an iCAT resource server, which uses a relational database to organize the content of the zone and to maintain iRODS metadata. The iCAT server stores metadata in the form of "triples" in its relational database. The triples consist of an attribute field, a value field, and a unit field. The content of each

[42] http://irods.org/wp-content/uploads/2012/04/iRODS-Overview-November-2014.pdf

of these fields can be independently defined and applied. Metadata may be user-defined or applied automatically.

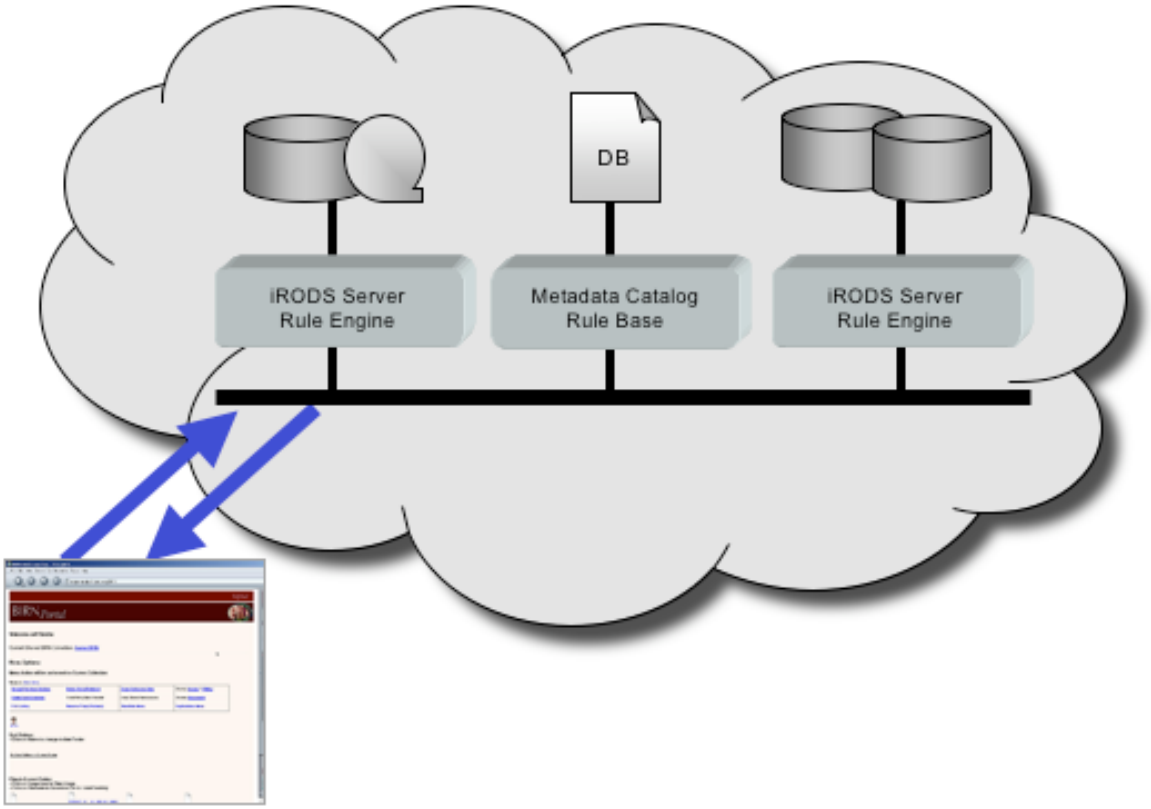

Figure 2 iRODS peer-to-peer architecture[43]

Once metadata is applied, it can be used in various ways. It can be used to trigger actions, based on rules defined in the iRODS rule engine. iRODS metadata can be searched as well. A simple way to search is using the iRODS meta command. More complex queries can be generated using a subset of SQL operations issued through the *iquest* command. One of often cited scalability limitations of iRODS is the centralized iCAT server, which is based on PostgreSQL database and holds the metadata.

6.3 Dynamic Federations

The Dynamic Federations[44] main goal is to connect geographically distributed storage sites. It creates a dynamic name space, consisting of meta-data items taken on demand from various endpoints. The Dynamic Federation solves the two main issues of distributed storage, composed of independent storage systems: dark data and dangling (outdated) references. The system can make use of static

[43] https://wiki.irods.org/index.php/iRODS_Architecture

[44] http://federation.desy.de/DynaFeds/The_Dynamic_Federations.html

file location catalogues, like the LFC[45], as hints for the location of the data. The performance has been optimized to federate storage endpoints or caches in a high speed, low-latency local area network, as well as to gap high latencies between different sites.

HTTP and WebDAV clients can browse the Dynamic Federation as if it were a unique partially cached name space, redirecting them to the appropriate endpoint for the actual data transfer. Standard mechanisms are available to provide all valid endpoints to the client, allowing it to download the data in parallel from all sources at the same time.

The typical use case is to present a huge distributed repository as if it were one, without the need of keeping an always up-to-date index of all the files it contains.

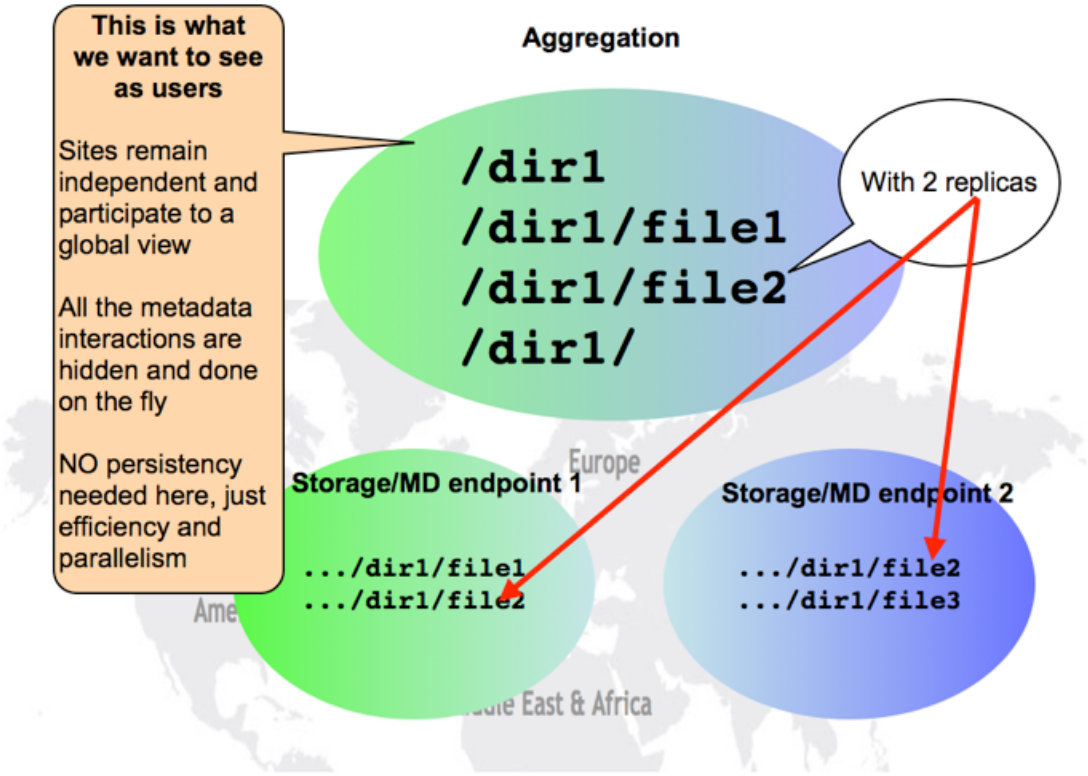

Figure 3 DynaFed namespace federation

The Dynamic Federation System is developed by the CERN Data Management team and deployed by CERN and DESY.

[45] https://twiki.cern.ch/twiki/bin/view/LCG/LfcAdminGuide

With respect to open data Dynamic Federations allow to provide a unified view over large data sets distributed across many storage sites, however the limiting factors include the support for only HTTP based WebDAV protocol without legacy POSIX access.

6.4 Globus Connect

Globus Connect[46] is a client-server solution allowing users and researchers to use the Globus transfer service. It simplifies the way of creating Globus endpoints - the different locations where data can be moved to or from using the Globus service. It is free to install and use for users at non-profit research and education institutions.

Globus Connect comes in two versions:

- Globus Connect Personal is designed for use by a single user on a personal machine. It is available for Mac OS X, Windows, Android and Linux operating systems.
- Globus Connect Server is designed to be installed by a system administrator on multi-user computing and storage resources. It is available for all major Linux distributions and integrates with existing IT infrastructure.

Installing Globus Connect sets up a GridFTP server for use with Globus.

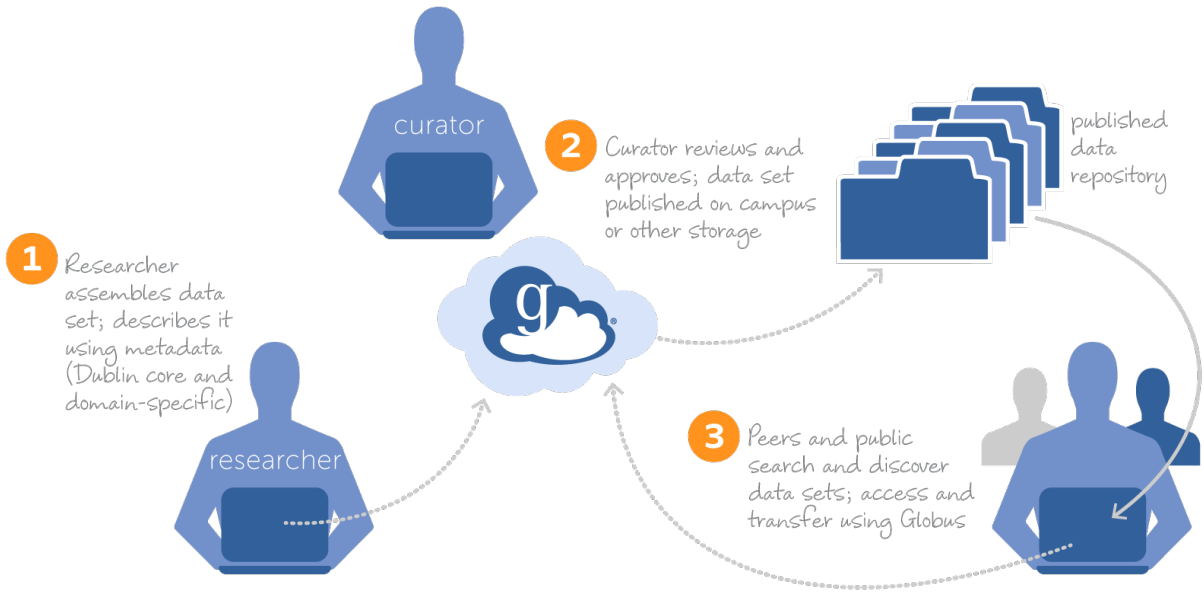

Figure 4 Globus Connect research data management flow

From the point of view of open data Globus Connect supports common protocols used in research institutions such as GridFTP as well as the integration of the GSI security infrastructure as well as sharing files with other Globus users. Currently Globus is enabling various features beyond simple transfer service, enabling open data access such as data sharing and data publishing. However, since

[46] https://www.globus.org/globus-connect

it is still mostly a transfer service based on GridFTP protocol, it does not allow POSIX read/write access to remote content.

6.5 Onedata

onedata[47] is a globally distributed storage solution, integrating storage services from various providers using possibly heterogeneous underlying technologies, such as Lustre, GPFS or other POSIX-compliant file systems and provides to clients interfaces based on CDMI, REST API and virtually mounted POSIX file system.

onedata has support for federated HPC applications, allowing transparent access to storage resources from multiple data centers simultaneously. onedata automatically detects whether data is available on local storage and can be accessed directly, or whether it has to be fetched from remote sites in real time.

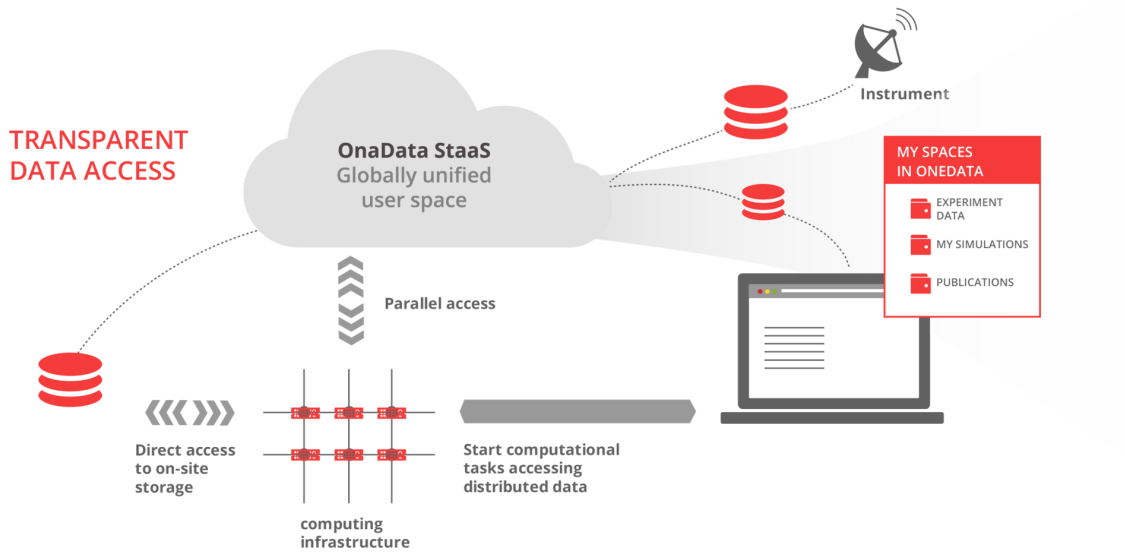

Figure 5 onedata overall vision

The core concept behind onedata system is Space, which can be considered as a virtual volume, which can contain regular files and folders, distributed across multiple data centers. Each user can create their own spaces, and share their content with other users, using customizable access rights, either *nix style or using complex Access Control Lists. This makes it easy to create ad-hoc collaborations between various users, without the need to involve administrators in the establishment of a Virtual Organization (e.g. using VOMS).

[47] https://onedata.org

onedata architecture comprises of 2 major components: oneprovider and oneclient. The former is installed within the data center and provides a unified interface to multiple file systems used in the center. Servers can scale to thousands of instances in order to improve performance. The client connects to the providers, which the user registered in the onedata portal, and its spaces are automatically provisioned by such providers. In the simplest case the user does not need to know the data stored with providers, although if necessary certain files can be pinned to certain locations.

Support for federation in onedata is achieved by the possibility of establishing a distributed provider registry, where various infrastructures can setup their own provider registry and build trust relationship between these instances, allowing users from various platforms to share their data transparently.

onedata provides an easy to use Graphical User Interface for managing storage Spaces, with customizable access control rights on entire data sets or single files to particular users or groups.

With respect to open data, one of the key features of onedata is the support for accessing and exchanging data across different infrastructures in a federated manner.

6.6 Technology feature comparison

The table below presents the comparison of specific technological features relevant for open data access between the analysed technologies.

	ownCloud	iRODS	Dynamic Federations	Globus Connect	onedata
Cloud based data access (CDMI)	-	In development (read only front end)	-	-	x
Legacy data access (POSIX)	-	x	-	-	x
Replica location	-	-	-	x	x
Data replication	-	x	x	x	x
Transparent replicas[48]	-	-	x	x	x
Metadata management	-	x	x	-	x
ACL management	x	x	x	-	x
Data sharing	x	-	-	x	x

[48] Regular users do not need to be aware of any replicas, they are managed by the data management system and presented to the user in a transparent way. Advanced users may whish however to access specific information about replica locations.

Data sharing for anonymous users	x	-	-	-	x
Data movement API	x	x	-	x	x

The following table presents how the requirements identified in the communities are supported by the analysed technologies.

Requirements	**ownCloud**	**iRODS**	**Dynamic Federations**	**Globus Connect**	**onedata**
REQ1: Publication of data based on certain policies	-	x	-	-	Planned (EGI-Engage)
REQ2: Make large data sets available without transferring them completely	-	x	x	-	x
REQ3: Complex metadata queries	-	x	-	x	x
REQ4: Integration of the open data access data management with community portals	-	-	-	-	Planned (EGI-Engage)
REQ5: Data identification, linking and citation	x	-	-	x	x (DOI to be implemented)
REQ6: Enable sharing of data between researchers under certain conditions	x	x	-	x	x
REQ7: Sharing and accessing data across federations	-	-	x	-	x
REQ8: Long term data preservation	-	x	-	x	Ongoing integration with EUDAT
REQ9: Data provenance	-	-	-	-	Planned (INDIGO DataCloud)

Based on the above comparison it is visible that requirements from the communities are highly heterogeneous and no single solution supports them all out of the box. Among the reviewed solutions onedata, which was planned as the technology to be used for the Open Data Platform prototype, has most of the required features and has clear development plans to cover missing functionalities in already funded projects such as INDIGO DataCloud[49] and EGI-Engage.

It is important to note that from an architectural point of view, onedata will not replace the communities' existing data management solutions, to which they are accustomed, but will act as a proxy between these platforms and Open Data interfaces provided to third party users interested in accessing the open data managed by EGI. This will involve development of GridFTP and iRODS support in onedata, which is envisioned in the framework of EGI-Engage. Additionally, no changes will be required at communities' existing data management solutions – the interfaces will be in front of One Data.

[49] http://cordis.europa.eu/project/rcn/194882_en.html

7 Implementation plans

This section presents the recommendations for open access data management platform, which will be developed within EGI-Engage.

7.1 Open Access Data architecture vision

The Open Access Data platform in EGI will enable management of open access data contributed by various EGI user communities, taking into account their specificities. The open data platform prototype will be based on onedata technology, and will act as a gateway between EGI Resource Centres and external users and services such as OpenAIRE, citation indexes. The platform will manage data migration, provision of persistent links to data objects, optimization of access and enable sharing of data between researchers and research groups (see Figure 6).

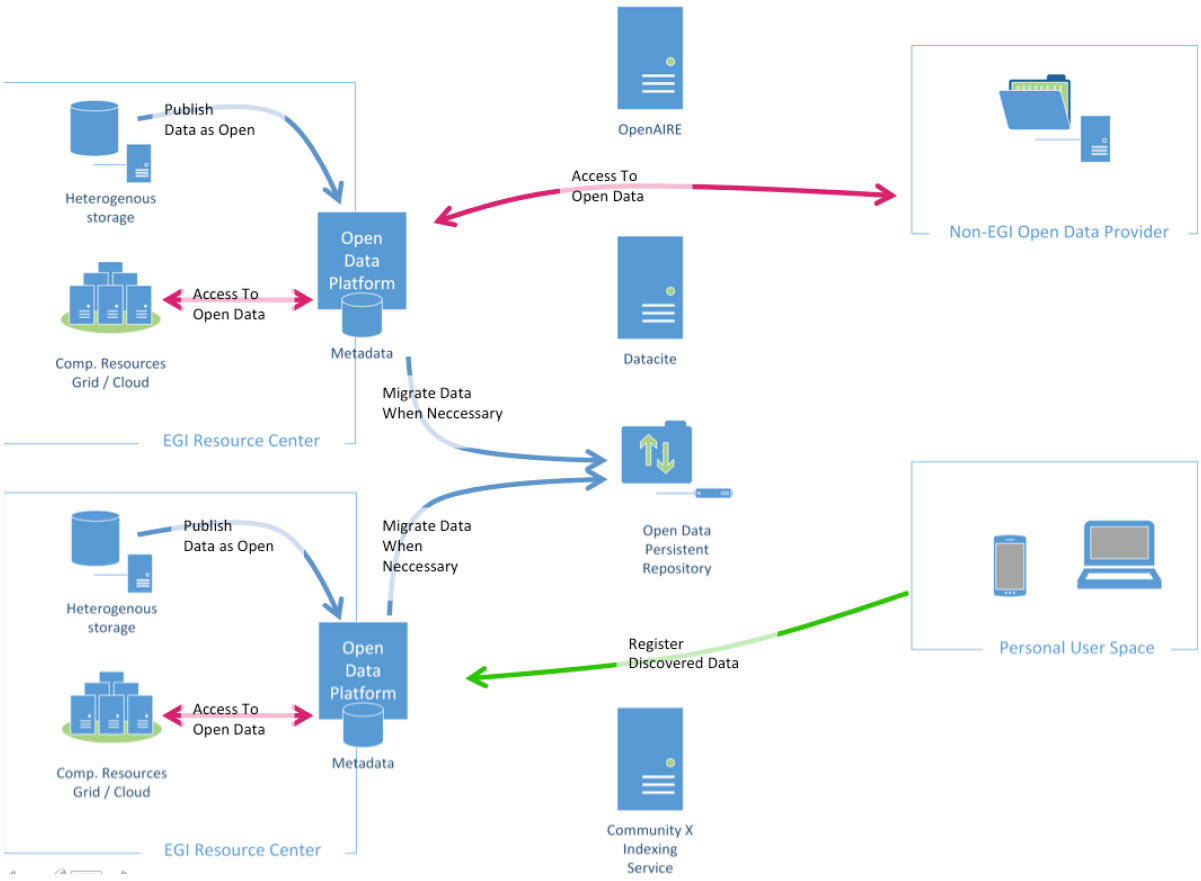

Figure 6 Open Access Data platform architecture

The Open Data Platform will function as a proxy between EGI based infrastructures and other data providers and users. Users from outside of EGI will be able to access data stored on various storage

solutions used in EGI based resource centers. Furthermore, users will be able to register published data through the Open Data Platform on EGI resources. EGI users will be able to directly publish data registered in EGI resources centers and Open Data Platform will handle all data requests coming from outside of EGI infrastructures. Open Data Platform will furthermore handle access to open data from computational resources in EGI sites.

As it is presented in Figure 6 open data platform works as a gateway in both ways: it connects open data existing outside of EGI infrastructure and enables it for data processing under EGI infrastructure without pre-staging data sets; it opens the data stored in the EGI platform and makes them open for other clients – not necessarily ones working under EGI federation. In the second case, open data platform is not only gateway for the actual data but it manages process of distributing metadata about published data to specialized services in open data discovery. It is important to underline that there are many services helping in data discovery, however for the pilot platform we are focusing on one selected service, OpenAIRE, which seems to be flexible enough to work together with our pilot platform. The distribution of metadata will be based on the OAI-PMH protocol.

In case of discovering open data sets by external services (e.g. OpenAIRE), users will be able to register the discovered data as the virtual data sets, which could be shared between their local communities if necessary, but more importantly users will get transparent access to the data sets with some optimization algorithms (like adaptive caching) improving processing performance.

When users decide to publish their data sets as public and for instance generate DOI identifies, it might be important to provide some sort of long-term data preservation service. In that case the data set might need to be migrated to some other storage services specializing in data preservation (e.g. EUDAT). In that case open data platform on behalf of the data owner might migrate the data for long-term preservation. However even though the process might take some time, the data will be available right away while the migration process is performed in the background. Migration to the other services will be optional, but even in such case the user will have control of the data, and the long term preservation copies will be seen in the open data system as another data replica.

In Figure 6 Open Data Platform is deployed on the selected EGI resource centres. In practice this is only needed in centres needing to provide services supporting data processing with higher speed. Thanks to that some local cache will be possible. In case that remote access to data is enough, clients can connect to their data using CDMI or POSIX virtual file system even when the particular resource centre is not hosting Open Data Platform.

7.2 Gaps between Requirements and Technologies

Based on the collected requirements and analysed technologies, the previously considered for open access data platform solution, onedata, seems the most feasible solution for building the prototype. This is due to inherent support for such features as:

- Support for federated data management,
- Provision of direct access to remote data sets over legacy POSIX protocol without need for migration,
- Easy sharing of data sets between users through concept of Spaces,
- Support for advanced metadata searches based on CDMI API's.

However, several gaps have to be fulfilled and developed within EGI-Engage in order to support wider set of communities, including:

- Rules for automatic publication of data sets based on certain rules (e.g. time since creation) or easy support for enabling such features in the communities user interfaces **(Milestone M20),**
- Identification of data objects using global identifiers such as DOI (Data Object Identifier) **(Milestone M20)**,
- More flexible approach to metadata and complex querying using various metadata standards **(Milestone M29)**,
- Enabling integration with community portals for open data information harvesting like OpenAIRE[50] in order to enable discovery of open data by OpenAIRE stored and maintained by the Open Data Platform, as well as the import of OpenAIRE information in order to replicate data in Open Data Platform **(M29)**,
- Development of additional protocols for integrating community storage solutions:
    - GridFTP **(Milestone M20)**
    - Dropbox **(Milestone M20)**
    - iRODS **(Milestone M29)**

7.3 Recommendations on Priorities for Developments

The following priorities for further development of Open Data Platform are proposed:

- Selection of pilot communities - LifeWatch seems to be a good candidate for preliminary testing. At the time of writing this document, work is already ongoing in integrating the LifeWatch data repositories with onedata
- Deployment of onedata as an EGI service,
- Implementation of missing functionalities in order to perform a cycle of data management and publication to a selected open data portal,
- Implementation of missing functionalities to perform a cycle of accessing and processing open data on the EGI infrastructure, open data coming from external to EGI repositories and sources,

[50] https://www.openaire.eu

- Implementation of additional protocols in order to integrate onedata with legacy community storage solutions in terms of data transfers and discovery. In this case onedata will act as a proxy providing unified interface for open data for communities using either onedata or other storage solutions, including custom platforms developed within communities.

7.4 Implementation time plan

The implementation plan for Open Data Platform based on onedata follows the general milestone schedule of EGI-Engage:

- **M20** – The first prototype version of the Open Data platform software stack ready to be deployed on the pilot sites and providing the minimal set of the planned functionalities to make it valuable in the data processing chain. This prototype will include GridFTP and Dropbox protocols for integrating community storage solutions. This milestone will be achieved within the following subtasks:
  - JRA2.1.1: Analysis of open data use cases and requirements
  - JRA2.1.2: Design and develop the Open Data platform prototype

  Although DropBox was not explicitly mentioned during the requirement gathering, it is felt that interoperability with DropBox is highly desired in the light of the increasing popularity of sync-and-share services.

- **M29** - This deliverable will be both demonstrator of the final version of prototype deployed on the selected pilot sites and an experience report based on the 6 months of evolution of Open Data prototype since the first prototype is delivered. This report will provide a summary about the scientific use cases supported in the infrastructure with involvement of open data and potential future work and functionalities related to open data. This milestone will additionally provide integration with iRODS in order to support EUDAT based storage infrastructure. This milestone will be achieved within the following subtasks:
  - JRA2.1.2: Design and develop the Open Data platform prototype
  - JRA2.1.3: Open Data platform demonstrator

8 Conclusions and future work

This report presented the results of a comprehensive requirements collection among various user communities related to EGI-Engage from such areas as biological and medical sciences, environmental and Earth sciences, agriculture as well as astronomy and astrophysics.

For the purpose of the requirements collection, a custom template has been prepared, focusing on issues related to open data access within the communities and identification of their current data management issues and solutions.

Based on the detailed requirements questionnaires (https://documents.egi.eu/document/2546), a summary table focusing on the most important aspects related to the open data access issues has been prepared and presented in the document.

Furthermore, an analysis of state of the art technologies potentially enabling open data access has been performed.

Based on the analysis, it was concluded, that as was planned in the EGI-Engage proposal, onedata platform should be used as the basis for open data access solution developed in the project. However, several new features have to be developed in order to support as wide number of community use cases as possible.