EGI-Engage

Open Data Platform: Requirements and Implementation Plans

M4.1

Date	31 Aug 2015
Activity	WP4
Lead Partner	CYFRONET
Document Status	FINAL
Document Link	https://documents.egi.eu/document/2547

Abstract

This report includes requirements from selected communities interested in 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 overview of data management technologies related to open data provision. Identification of technological gaps with respect to the requirements has been created based on comparison between requirements and available technologies and the recommendation for technology selection and development priorities has been 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:	Sandro Fiore Giacinto Donvito	UNISALENTO INFN	10 Aug 2015
Approved by:	AMB		31 Aug 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, T. Ferrari
V.2	6 Aug 2015	First version for external review	B. Kryza, L. Dutka, Y. Chen, T. Ferrari
FINAL	31 Aug 2015	Final version	B. Kryza, L. Dutka, Y. Chen

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
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 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 has been prepared.

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 kilobytes to terabytes), 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, a set of significant common requirements has been 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, enable selective data sharing between researchers and research groups, long time data preservation and in some cases data provenance.

After analysis of existing technologies, it has been verified that for development of the Open Access Data platform in EGI Engage, onedata solution will be used. The most important features of onedata include easy way of sharing data between individuals and groups of users and federation support. However, several new features have to be developed in order to support as wide number of community use cases as possible.

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.

2.2 Open Data Access platform vision and goals

The main problem with current computing infrastructures in Europe is the lack of support for fostering data dissemination and exploitation through implementation of data management plans allowing data preservation, distribution and provision over extended periods of time to research communities outside of communities initially involved in the various infrastructure projects. The vision of Open Data Access platform for EGI, realized through EGI-Engage, is to prototype architecture and technology for realizing such open data vision.

Eventually, Open Data Access platform will extend the EGI Federated Cloud with data preservation and access features, including data caching and bringing the computation to where the data is located already.

Furthermore, through the SME engagement activities, the open data solution will support the mission of the European Big Data Value, whose mission is to "… *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*".

2.3 Our Problems

The main issues and challenges that have driven the preparation of this report included:

- How to better understand the communities requirements
- How to efficiently get desired information
- How to efficiently communicate and manage the complex process
- How to conduct valuable analysis and reveal insights
- How to provide useful recommendations for development

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 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. 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 public 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[1], 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;
- Focus on data, computation, and use of e-Infrastructures;
- Focus on EGI user communities, in particular EGI-Engage Competence Centers.

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 technology, and gives the recommendation of the priorities for developments. Finally, Section 8 concludes this work.

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

3 Methodology

This section provides overview and introduction into methodologies used for the process of requirements questionnaire design, collection and analysis.

3.1 Design and Use of Template

A Generic Template Design for EGI Engage Requirement Gathering Tasks

The requirement collection is a challenging task. In order to gather requirements from user communities in a systematic way, we have designed a generic template[2] for EGI Engage project for requirements gathering from various communities.

The generic template provides a structured framework with guiding questions. It captures the state-of-the-art experiences from various EGI involved projects, such as INDIGO, EGI-InSPIRE, EGI-Engage, and ENVRI. It is based on the Open Distributed Processing (ODP) framework, an ISO standard, and uses a case-study driven approach. A **Case Study** is an implementation of a research method involving an up-close, in-depth, and detailed examination of a subject of study (the case), as well as its related contextual conditions. The Case Study will be based on a set of **User Stories**, i.e. how the researcher describes the steps to solve each part of the problem addressed. ***In practice, the user community shall be notified that the selection of the use stories shall be representative reflecting both of the research challenge and complexity, and of the possible solutions offered by the investigation project***. User Stories are the starting point of **Use Cases**, where they are transformed into a description using software engineering terms (like the actors, scenario, preconditions, etc. Use Cases are useful to capture the requirements that will be handled by the technology provider, and can be tracked, e.g., by a Backlog system from an Open Project tool[3]

A case study is built incrementally by interacting with the users' overtime. The complete description of the case study shall picture different aspects of the system required, including sufficient information for future analysis or implementations. Using ODP framework, the template is designed to examine the requirements for a system from 5 different aspects:

- **The Science Viewpoint**, concerns the organisational situation in which the research activity in the current case is to take place.
- **The Information Viewpoint**, concerns modelling of the shared information manipulated within the system of interest.
- **The Computational Viewpoint**, concerns the design of the analytical, modelling and simulation processes and applications provided by the system.

[2] EGI Engage generic Template for requirement collection https://wiki.egi.eu/wiki/Requirement_Collection
[3] Open Project tool: https://www.openproject.org/

- **The Engineering Viewpoint**, tackles the problems of diversity in infrastructure provision; it gives the prescriptions for supporting the necessary abstract computational interactions in a range of different concrete situations.
- **The Technology Viewpoint**, which concerns real-world constraints (such as restrictions on the facilities and technologies available to implement the system), applied to the existing computing platforms on which the computational processes must execute.

The design of the template also considers the following aspects:

- **Functional and non-functional requirements**. Apart from functionalities, non-functional aspects shall be inquired, which includes, e.g., performance, privacy issues, etc.
- **Current situation and requirements for a future system**. In many situations, a user community couldn't provide the precise description of requirements for a future system. This maybe because the community/community contacting people couldn't assess the new technology to be enabled by the development team at that time. However, information about current system is still useful for analysing their needs. The template provides areas for the descriptions of both current system and the requirements for a future system.
- **Structured questions and flexibility for extension**. Many sections provide structured questions, which are based on EGI experiences and other state-of-the-arts. The intension is to capture the existing experiences and provide a knowledgebase where a requirement collector can refer to when preparing the questionnaires or interviews. There are also spaces/fields for "free-hands" inputs, considering new issues/topics may arise from inquired communities.
- **Mandatory and optional input fields**. Mandatory fields are marked by bold text, which are highly recommended to be filled in order to have sufficient information. When all mandatory fields are filled, the requirements collection can be treated as completed.
- **Review and approval**. The information gathered shall be reviewed internally, and approvals from inquired communities shall be obtained in order to validate the preciseness of the contents.
- **Status of the information collection**. Information may be gathered over time, the status of the requirement collection shall be documented.

The instruction of using the template is given at EGI wiki site[4].

The template can be used in various purposes, for example:

- Can be used to extract relevant requirement information from community design documents, website, and presentations;
- Can be used as a recording form during requirement interview meetings;
- Can be used as questionnaires being sent to user communities to collect information;
- Can be used to organise information incrementally gathered from different sources, emails, and conversations with different people in different contexts.

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

The benefits of using the template include, but not limited to:

- To help a requirements collection team better scope the investigation and plan the activities. For example, in the first section of template, the scopes and purposes for the requirement collection shall be filled at the initial stages. With the help of the technology development team, key technology issues concerned by the development team shall be identified. Based on these, the generic template shall be customised to be more suitable for the specific requirement collection scope and purposes, e.g., remove sections or questions not essential, and add specific questions that may help to drill down into the details of the interested areas. Space for planning of the activities is given, where a series of activities can be organised, such as, preparation of the template, reviewing of the questions, gathering information, interviews of community representatives, etc.
- To improve the communications efficiency within internal team, between communities, and between technology development team.
- To help managing the requirement gathering processes in an efficient way which can ensure the quality of the work. For example, the template status of the information collection can be recorded, thus tracked. The template defines the following status
  - **PENDING**: Requirement gatherers have been identified but have yet to start work.
  - **GATHERING**: Information about the requirement is being gathered and recorded.
  - **COMPLETE**: Gathering / recording information about the requirement has been completed.
  - **REVIEWING**: The information is being reviewed and cleaned up, internally by the team.
  - **CONFIRMING**: Information about the requirement is being reviewed / confirmed by communities and experts. (The name of such a person shall be provided at the end of each session indicated field).
  - **ACCEPTED**: Information about the requirement is complete, accurate and accepted as correct by all stakeholders.
  - **STOPPED**: Work on this topic has been interrupted for the reason specified

  Moreover, in order to ensure the information collected is valid and up-to-date, the template requests the approvals from relevant peoples to be obtained.

Using Template to Collect Requirements for the Open Data Platform

Following the guidance given by the generic template (refer to the EGI wiki site), 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 firstly meet with the technology development team to understand what aspects that Open Data Platform would want to inquire the community.

Prepare the template: Open Data Platform would like to identify the current requirements, challenges and expectations of the communities interested in making their data public within the EGI framework. Based on the technical expectations, the questions to inquire communities are focused on:

- What kind of data, in what formats and sizes is managed by the community?
- What are the life cycles of data created within the community?
- What are the current data management and transfer technologies 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 focus on the EGI user communities, in particularly, those participated in EGI Engage Competence Centres and those have been involved in the discussions of the Open Data Platform, for example, we organised Towards Open Data Cloud session[5] in the EGI User Forum 2015, Lisbon, 19-22 May, where we invited various EGI user communities to give talks and discuss the issue. We use Google sheet to maintain a list of communities for inquiry and interview purpose.

Collect requirements: with the limited resources and time constrains, we decide that the requirements collection team to first extract desired information based on available material, e.g., communities' design documents, presentations, web and wiki sites, and various documentations, then send it to communities to clarify and to obtain missing information.

Review and approval: to ensure the quality of the information collection, we enforce that the collected information shall be reviewed by internal team and shall be approved by communities.

[5] 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

4 Research Communities and Their Use Cases

This section presents an overview of the research communities considered in the requirements collection, grouped into 3 categories: biological and medical sciences, environmental and Earth sciences, agricultural sciences and astronomy and astrophysics.

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 storage 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.

4.1.2 MoBRAIN and Structure biology

The main objective of the MoBRAIN Competence Centre is to lower barriers for scientists to access modern e-Science solutions from micro to macro scales. MoBRAIN builds on grid- and cloud-based infrastructures and on the existing expertise available in WeNMR[6], N4U[7] and technology providers (NGIs and other institutions, OSG). This initiative aims to serve its user communities, related ESFRI projects (e.g. INSTRUCT) and, in the long term, the Human Brain Project (FET Flagship), and strengthen the EGI services offering.

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.

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

[7] https://neugrid4you.eu

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 CC 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 requires that certain security concerns are satisfied (data is not opened to the public by default) and metadata about how the data was sequenced is available.

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 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)*

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

[9] From http://www.emso-eu.org

4.2.2 LifeWatch

LifeWatch is a part of the European Strategy Forum on Research Infrastructure (ESFRI) and can be seen as a virtual laboratory 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.

LifeWatch infrastructure supported through EGI-Engage is designed around the collection of data from sensor networks. Data collected from sensor networks is transferred to the processing network based on Federated Cloud LifeWatch infrastructure. 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.

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[10] project, supported by the Agriculture Information Management Standards of the Food and Agriculture Organization of the United Nations (AIMS FAO)[11] and the CIARD[12] global initiative, introduces a set of recommendations applying to agri-food research community for 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[13] mandate and share their metadata with their thematic aggregator in order to publish them in OpenAire[14]. [15]

[10] http://aginfra.eu

[11] http://aims.fao.org

[12] http://www.ciard.info

[13] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

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

[15] From www.aginfra.eu

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. According to the focus group participants examples of open access sharing of research data are difficult to spot in the domain of agricultural research. While open access document repositories only have to promote self-archiving of already published works, research data/datasets are assets most researchers are not willing to share openly.

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 on an unprecedented scale. It will allow significant extension of our current knowledge in high-energy astrophysics. The aims of the CTA can be roughly 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 access to RAW data to perform reduction tasks and then needs to store the reduced higher level data in the archive. Depending on the data policy, after 1 year all proprietary data will become public, so it will be available through the data access service also who are not Principal Investigators. After this automatic publication it will be possible to publish higher-level products to the Virtual Observatory servers in order to be directly usable by all VO tools available to the astronomical community.

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 major advantage of the phased array concept. It is not only used to increase observational efficiency, but may be vital 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.

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 at a set of data centres.[16] The main objectives of the community 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, 2 main use cases have been identified. First, all user data is made public after 1 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 portal and search through using FITS metadata contained in observation files and indexed in external databases.

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

5 Requirements Analysis and Findings

The goal of this section is to give a summary of the selected communities' requirements related to open data access as well as 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.

5.1.1 Open access policies[17]

Community name	Publication policy?	Is usage tracking required?	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	Tier 1 - Restricted use: Data developing analysis algorithms Tier 2 - Restricted Use: Data available non-conflicting research questions Tier 3 - Restricted use: Full data available for collaborative investigation, joint research questions
MoBRAIN				Most data is private to investigators, however it is planned that simulation data could be opened.
BBMRI				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

[17] Shaded cells mean, that no input has been provided yet

				Protection of Personal Data.
EMSO				
LifeWatch			Long-term	For managers IPA system at IFCA (similar to LDAP). Users TBD.
Agrodat.hu				
agINFRA	Data is free from beginning		Long-term	
CTA				
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			

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—will grow to O(1000PB) within next 5-10 years	Brain scans are stored in a form of: series of bitmaps, VTK (for 3d rendering), HDF5, 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 data (several 100MB) Analysed NMR data (several GBs)		PDB (Protein Data Bank), Text files		
BBMRI					

EMSO					
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, CYZ, others) Images/Videos stored in JPG	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					
agINFRA	~10KB	~1PB	XML, MCPD (Germplasm data)	Custom solution	HTTP
CTA		>1000PB (target size)	FITS, RAW, ROOT, JSON, XML, BSON		
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		
BBMRI	ICD-9, ICD-10, SNOMED CT, UMLS	
EMSO		
LifeWatch	Ecological Metadata Language (EML)	
Agrodat.hu		
agINFRA	RDF, OWL, XML, SKOS, OAI-PMH	Files, RDF Triple stores
CTA		
LoFAR		
CANFAR	FITS	

5.1.4 Key findings from the questionnaires

The analysis of 10 communities realized within EGI Engage showed several main issues:

- High heterogeneity of data management patterns
- Usage of various data formats and metadata standards
- Sometimes lack of clear data management policies (i.e. which and when data should be made public)

Since input from several communities on all or selected aspects of open data management in their use cases has not been available at the time of writing of this document, the requirements will be updated after more requirements have been gathered.

5.2 Identification of the Common Requirements

This section presents an attempt at 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 data based on certain conditions

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.

5.2.2 REQ2: Make large data sets available without migrating them

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: 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 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 communities 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 develop a more 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, and only where it is located.

5.2.6 REQ6: Enable 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 who 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 The 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 either obtained directly from experiments or generated through simulations.

6.1 ownCloud

ownCloud[18] is an open-source framework for creating self-managed file hosting services (Figure 1), similar to Dropbox. 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).

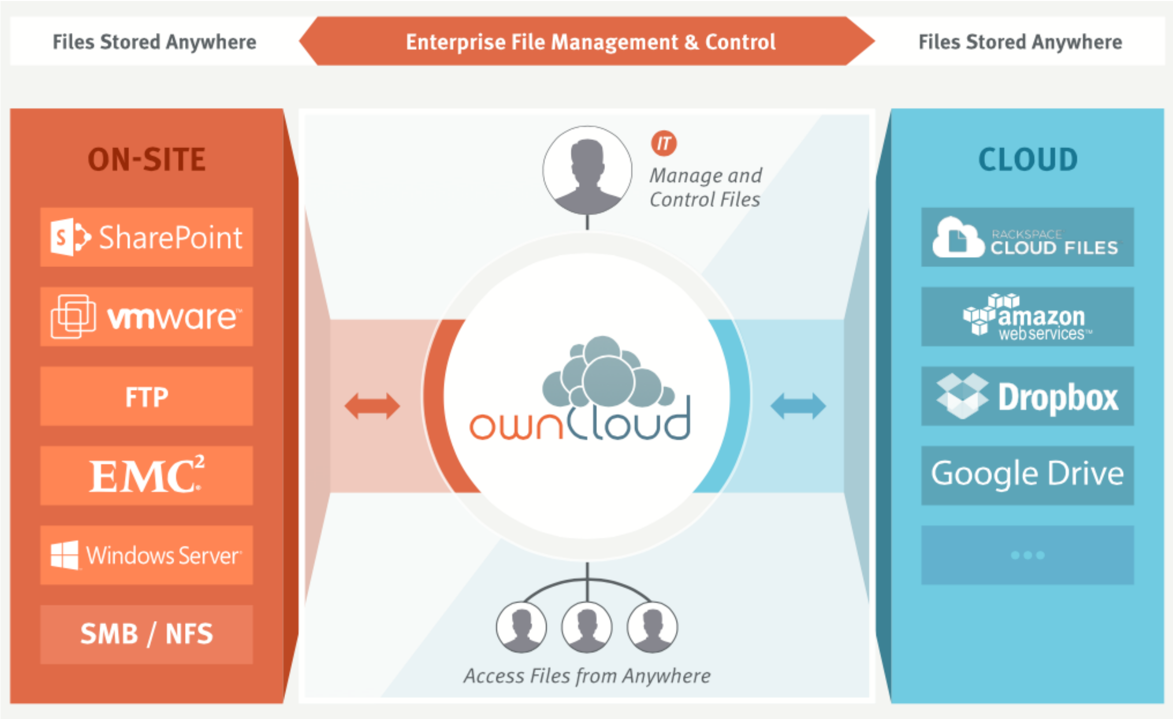

Figure 1 ownCloud overall functionality

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

From the point of view of open data, ownCloud supports publication of links do data sets (files) using public URLs. However, ownCloud is more focused on consumer applications, i.e. no support for HPC in terms of optimized file transfers or remote read/write POSIX access are available.

6.2 iRODS

The Integrated Rule-Oriented Data System (iRODS)[19] 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 of these fields can be independently defined and applied. Metadata may be user-defined or applied automatically.

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

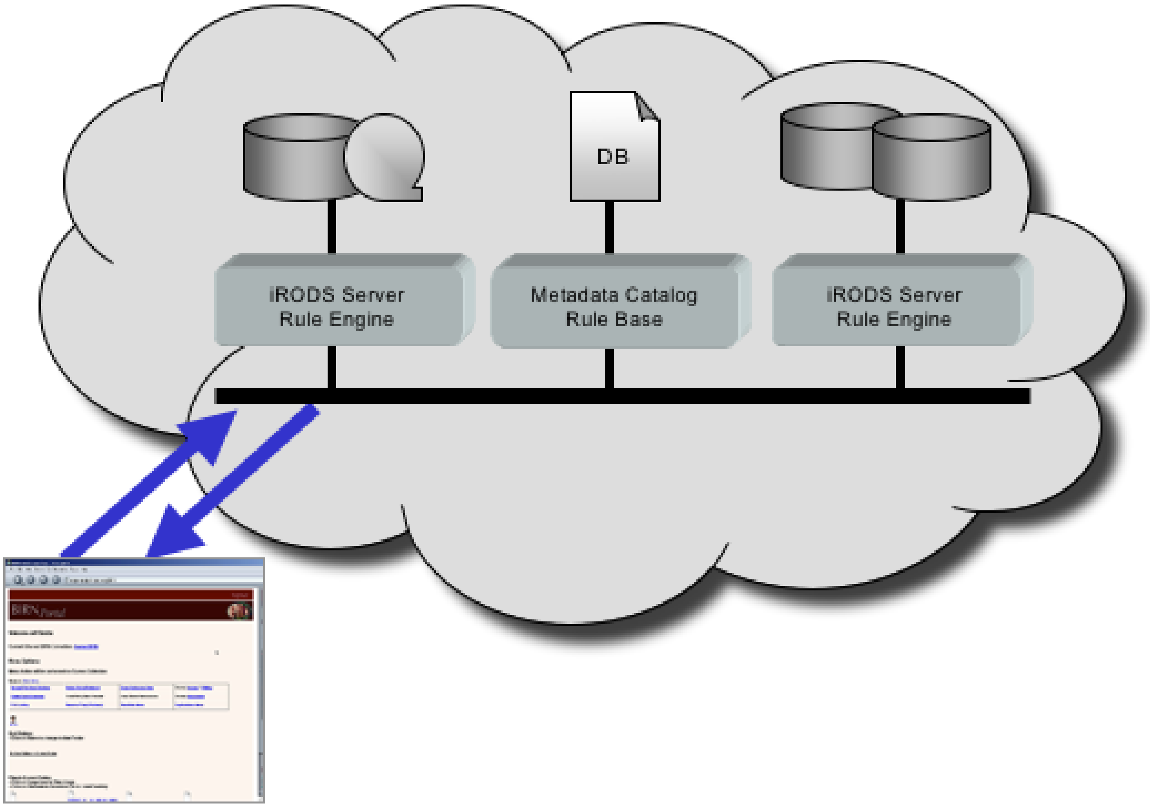

Figure 2 iRODS peer-to-peer architecture[20]

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 imeta command. More complex queries can be generated using a subset of SQL operations issued through the *iquest* command.

6.3 Dynamic Federations

The Dynamic Federations[21] 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 file location catalogues, like the LFC, 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.

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

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

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.

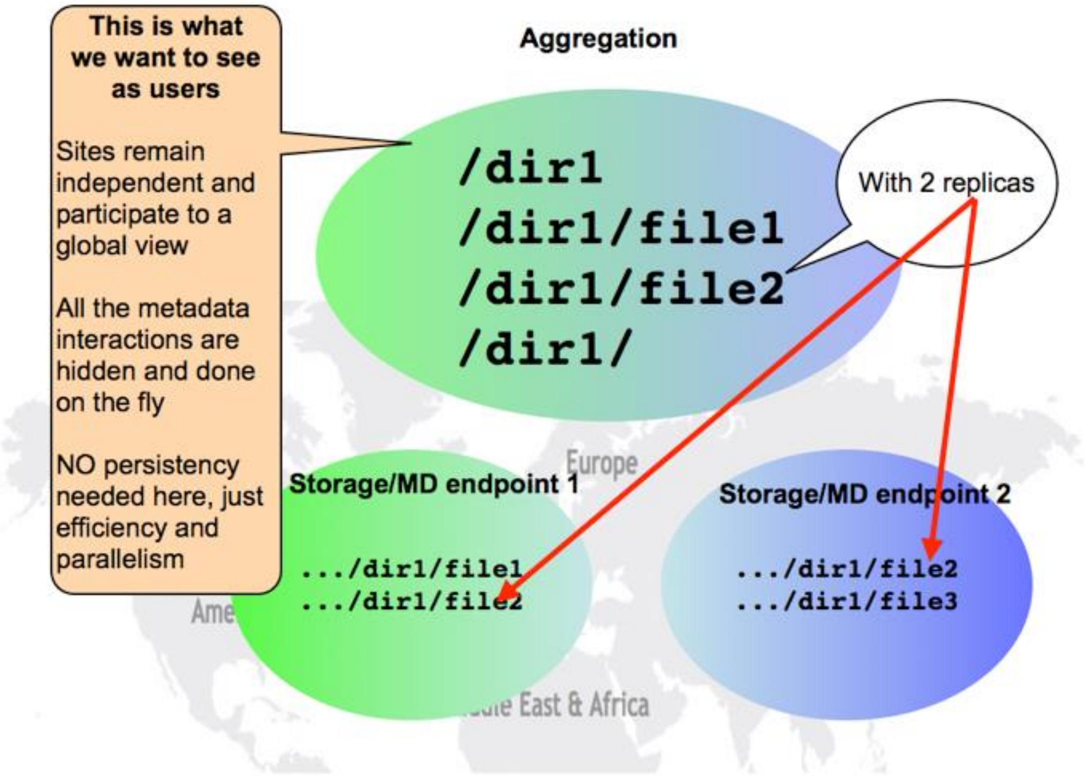
This is what we want to see as users
Sites remain independent and participate to a global view
All the metadata interactions are hidden and done on the fly
NO persistency needed here, just efficiency and parallelism
Aggregation
/dir1
/dir1/file1
/dir1/file2
/dir1/
With 2 replicas
Storage/MD endpoint 1
.../dir1/file1
.../dir1/file2
Storage/MD endpoint 2
.../dir1/file2
.../dir1/file3
Europe
Ame
East & Africa

Figure 3 DynaFed namespace federation

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

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 could be the support for only HTTP based WebDAV protocol without legacy POSIX access.

6.4 Globus Connect

Globus Connect[22] 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, 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.

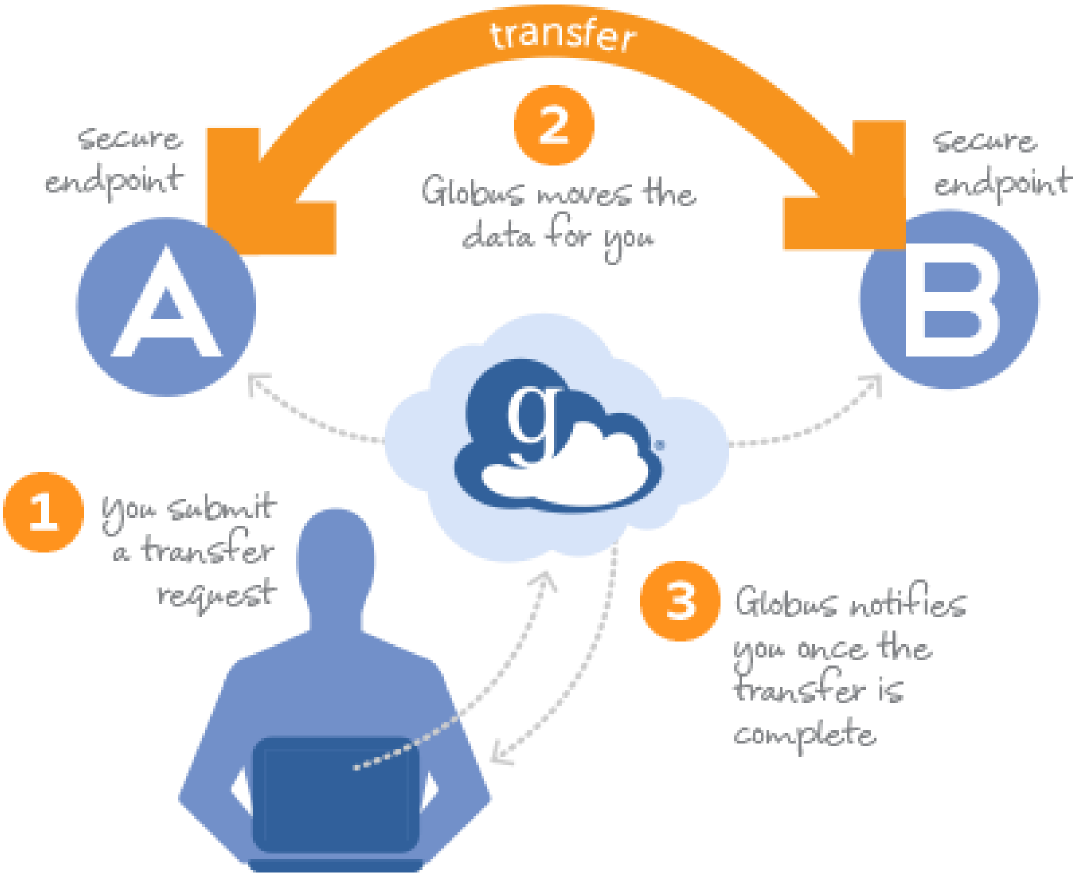
transfer
secure endpoint
A
2
Globus moves the data for you
secure endpoint
B
g
1
you submit a transfer request
3
Globus notifies you once the transfer is complete

Figure 4 Globus Connect data management flow

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

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. However it does not allow POSIX read/write access to remote content.

6.5 Onedata

onedata[23] 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.

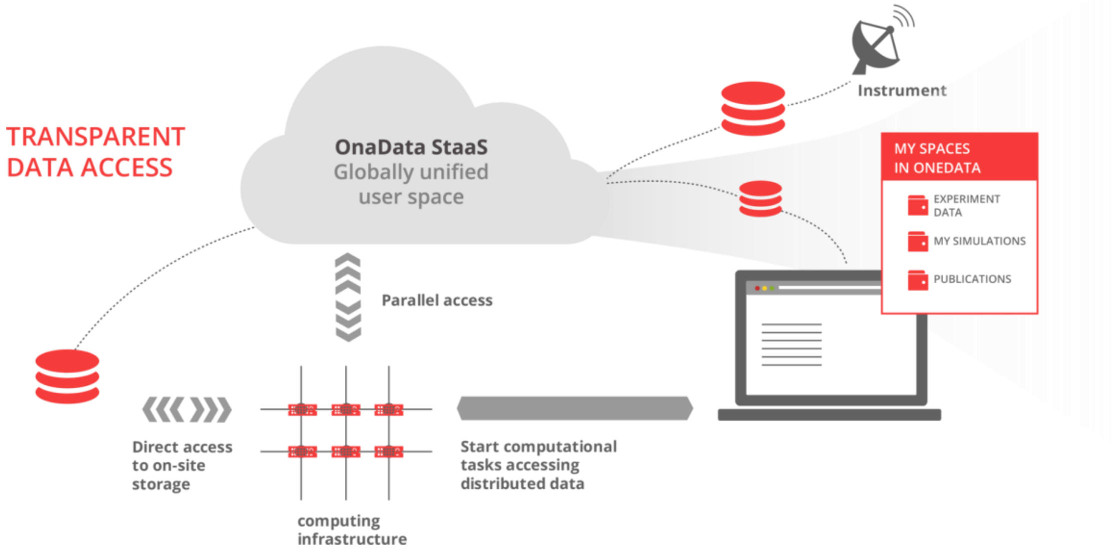

Figure 5 onedata overall vision

The core concept behind onedata system is Space, which can be considered as a virtual volume, which is 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).

[23] https://onedata.org

onedata architecture comprises of 2 major components: oneprovider and oneclient. The former is installed within 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 onedata portal, and his spaces are automatically provisioned from these providers. In the simplest case the user has no need to know which data is stored with which provider, 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.

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

Open Access Data platform in EGI will enable management of open access data contributed by various EGI user communities, taking into account their specificities. Open data platform 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).

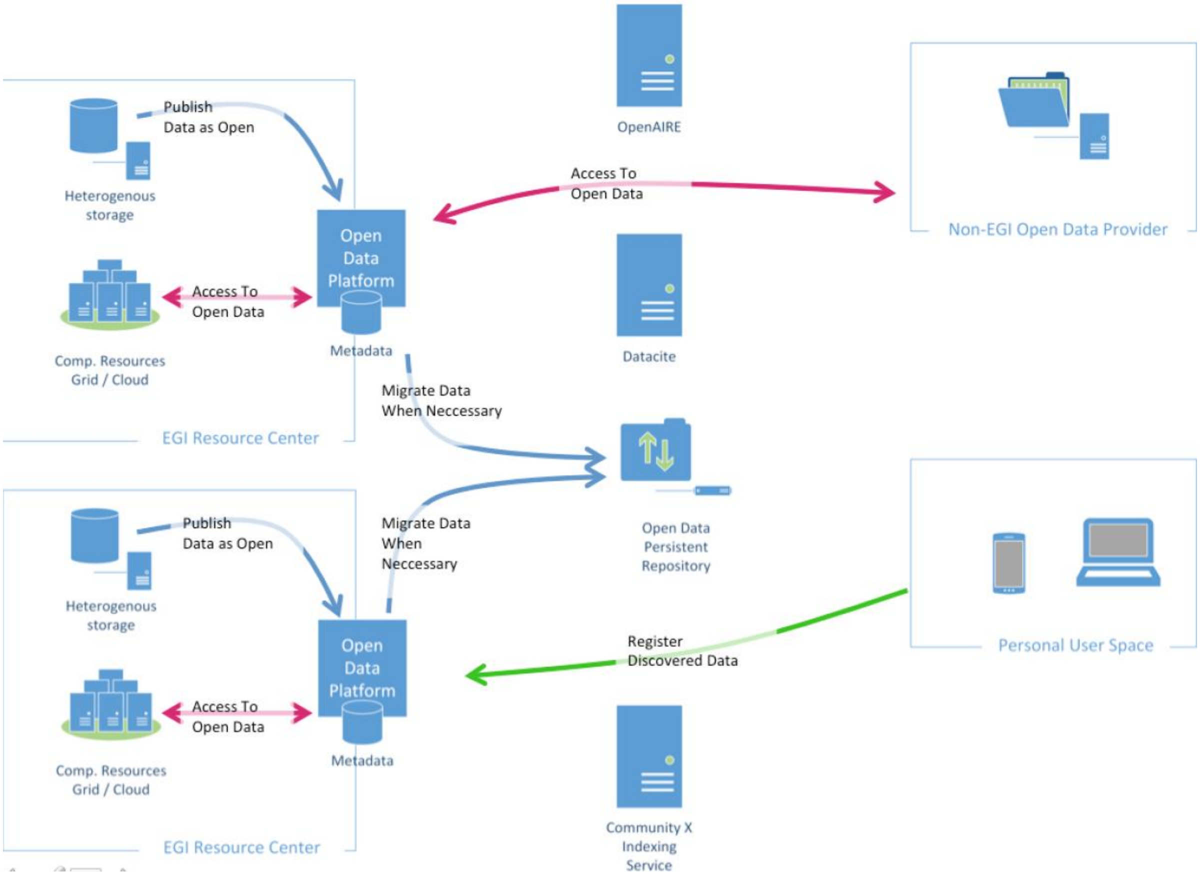

Figure 6 Open Access Data platform architecture

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 most feasible. This due to inherent support for such features as:

- Support for data 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
- Integration with open data portal like OpenAIRE[24] in order to automatically register open data stored and maintained by Open Data Platform.

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,
- Identification of data objects using global identifiers such as DOI (Data Object Identifier),
- More flexible approach to metadata and complex querying using various metadata standards,
- Enabling integration with community portals for data access.

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,
- 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.

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

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 (for which links are available in Appendix 1), 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.

Appendix I. Requirement Collections

A.1 Human Brain Project

The questionnaire is available at:

https://documents.egi.eu/secure/RetrieveFile?docid=2546&version=1&filename=Requirement%20Extraction_Open%20Data%20Platform_HBP_230715_v2.docx

A.2 MoBRAIN

The questionnaire is available at:

https://documents.egi.eu/secure/RetrieveFile?docid=2546&version=1&filename=Requirement%20Collection%20Template_Open%20Data%20Cloud_MOBRAIN_v3.docx

A.3 BBMRI

The questionnaire is available at:

https://documents.egi.eu/secure/RetrieveFile?docid=2546&version=3&filename=Requirement%20Collection%20Template_Open%20Data%20Cloud_BBMRI_v1.docx

A.4 EMSO

The questionnaire is available at:

https://documents.egi.eu/secure/RetrieveFile?docid=2546&version=1&filename=Requirement%20Collection%20Template_Open%20Data%20Cloud_EMSO_v2.docx

A.5 LifeWatch

The questionnaire is available at:

https://documents.egi.eu/secure/RetrieveFile?docid=2546&version=1&filename=Requirement%20Collection_Open%20Data%20Cloud_LifeWatch_v2.docx

A.6 Agrodat.hu

The questionnaire is available at:

https://documents.egi.eu/secure/RetrieveFile?docid=2546&version=1&filename=Requirement%20Collection%20Template_Open%20Data%20Cloud_AGRODAT_v2.docx

A.7 agINFRA

The questionnaire is available at:

https://documents.egi.eu/secure/RetrieveFile?docid=2546&version=1&filename=Requirement%20Collection%20Template_Open%20Data%20Cloud_agINFRA_v2.docx

A.8 CTA

The questionnaire is available at:

https://documents.egi.eu/secure/RetrieveFile?docid=2546&version=1&filename=Requirement%20Collection%20Template_Open%20Data%20Cloud_CTA_v2.docx

A.9 LoFAR

The questionnaire is available at:

https://documents.egi.eu/secure/RetrieveFile?docid=2546&version=1&filename=Requirement%20Extraction_Open%20Data%20Platform_LoFAR_v2.docx

A.10 CANFAR

The questionnaire is available at:
https://documents.egi.eu/secure/RetrieveFile?docid=2546&version=1&filename=Requirement%20Collection_Open%20Data%20Cloud_CANFAR_v1.docx