

Requirements Collection

for Open Data Cloud Platform

MoBRAIN

|  |  |
| --- | --- |
| **Author: Yin Chen** |  |
| **Version: v1.0** |  |
| **Document Link:** |  |

TABLE OF CONTENTS

Appendix Requirement Extraction Template 3

A.0 Purpose and Scope of the investigation 3

A.1 Science ViEWpoint 5

A.2 Information Viewpoint 7

A.3 TECHNOLOGY Viewpoint 9

# Appendix Requirement ExtractiOn Template

A.0 Purpose and Scope of the investigation

*This section is input by a requirement collector to explain the purpose and scope of the investigation to an inquiry community, explaining the instructions of how to fill the template, and to keep records of the status of the requirement collection progress.*

****A.0.1 Authors****

*All authors contributing***directly***to this focus. Incrementally add names here as people actually contribute.*

|  |  |  |  |
| --- | --- | --- | --- |
| **Roles** | **Contact Person** | **Organization** | **Contact email** |
| Project Leader | Tiziana Ferrari | EGI.eu | tiziana.ferrari@egi.eu |
| Technology Provider | Lukasz Dutka | Cyfronet | lukasz.dutka@cyfronet.pl |
| Requirement Collector | Bartosz Kryza | Cyfronet | bkryza@agh.edu.pl |
| Requirement Collector | Yin Chen | EGI.eu | yin.chen@egi.eu |
|  |  |  |  |

****A.0.2 Purpose and Scope****

|  |
| --- |
| **Purpose** *(Please describe the background, objectives and purpose of this requirement collection activities.)* |
| This requirement collection activity is organized within EGI-Engage project, aiming to support the development of Open Data platform. Based on this questionnaire Open Data Platform would like to identify the current requirements, challenges and expectations of the communities interested in making their data public within EGI framework. In particular the major aspects related to ODP that should be resolved through this questionnaire include:* 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 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.)
 |
| **Scope** *(By discussing with the technology provider teams, please briefly describe the technology to be provided, and intended inquiring areas)* |
| An Open Data Platform (ODP) will be designed to foster the discovery, dissemination and exploitation of open data in cloud environments, 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). ODP will be based on onedata data management solution (<http://www.onedata.org>). |
| Expectations(*By discussing with the technology provider teams, summarise any special expectations they would want to notify the requirement collection team)* |
|  |
| **Information approved by** | <Technology Provider> |

****A.0.3 Status of the requirement collection****

|  |  |  |  |
| --- | --- | --- | --- |
| **Description of the activities** | **Status** | **Responsible Person** | **Date** |
| Prepare the template | PENDING | Yin Chen, Bartosz Kryza | 10 Jul 2015 |
| Obtain approval of the template from the technology provider | PENDING | Lukasz Dutka | 13 Jul 2015 |
| Information collection from available materials  | GATHERING | Bartosz Kryza | 22 Jul 2015 |
| Review the information collected | REVIEWING | Yin Chen | 27 Jul 2015 |
| Send to the community for confirming and missing information  | CONFIRMING | Yin Chen | 27 Jul 2015 |
| Get approvals from the community | ACCEPTED |  |  |
| Complete information collection | COMPLETE |  |  |

* **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 filed).
* **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.

A.1 Science ViEWpoint

*Science viewpoint concerns community objectives to be achieved through the collaboration, and the details of use cases related to the technology to be provided. Information in this section needs helps and approvals from Research Managers of the user community.*

**A.1.1 Community Information**

|  |  |
| --- | --- |
| **Community Name** | A Competence Center to Serve Translational Research from Molecule to Brain |
| Community Short Name if any | MoBRAIN |
| Community Website | https://wiki.egi.eu/wiki/Competence\_centre\_MoBrain |
| **Community Description**  | 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](http://www.wenmr.eu/), [N4U](http://neugrid4you.eu/) 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. |
| **Community Objectives** | * lower barriers for scientists to access modern e-Science solutions from micro to macro scales
 |
| **Main Contact Institutions** | Bijvoet Center for Biomolecular Research Faculty of Science, Utrecht University |
| **Main Contact**(*name and email*) | Alexandre Bonvin <a.m.j.j.bonvin@uu.nl> |
| Prior requirement capture activities and ideally a summary and references to their outcome | <*input here*> |
| Upload copies of files and provide links to them | <*input here*> |
| Cite papers | <*input here*> |

**A.1.2 Collaborations with Open Data Cloud Project**

|  |
| --- |
| **Scientific challenges** (*Please describe your problems and motivations for the collaboration with* ***Open Data Cloud***) |
| * The sample is often the most precious part and is specific to one project and group, as such data are only made public upon publication, the only data sharing aspects may-be transferring/sharing data between experimental sites and the user lab(s)
* No standard policy and resources for raw data archival
* No clearly defined metadata standards yet
* No facility for permanent storage of raw data
* Derived data (structures, assignments, restraints) are permanently stored in public databases
* For integrity in research, data (including raw data) should ideally be stored for at least 10 years
 |
| **Objectives** (*Please describe your objectives to be achieved through collaboration with* ***Open Data Platform****)* |
| * Derived data shared in public databases
* Bring the cloud to the data
	+ Keep the data at data centers close to the experimental infrastructure
	+ Develop VMs for data processing at local data centers
 |
| Expectations *(please describe your expectations for the new technology to be provided by the* ***Open Data Platform****)* |
| <*input here*> |
| Impacts and Benefits (*Please be specific and use quantified indicators and targets wherever possible*) |
| <*Input here*> |
| *KPI inputs**(Please indicate as realistic as possible the expected results)* |
| *Area* | *Impact Description* | *KPI Values* |
| *Access* | *Increased access and usage of e-Infrastructures by scientific communities, simplifying the “embracing” of e-Science.*  | * *Number of users of the web portals: <input here>*
* *Number of sites provide the services:* <*input here*>
 |
| *Usability* | *Simplifying deployment of the web portals in cloud resources* | * *Number of downloads:* **<***input here***>**
 |
| *Impact on Policy* | *Policy impact depends on the successful generation and dissemination of relevant knowledge that can be used for policy formulation at the EU, or national level.*  | **<***input here***>** |
| *Visibility* | *Visibility of the project among scientists, technology providers and resource managers at high level.* | * *Number of citations of the software* **<***input here***>**
* *Number of portal cloud installations/usage:* **<***input here***>**
* *Advertisement at events/conferences/workshops:* **<***input here***>**
 |
| *Knowledge Impact* | *Knowledge impact creation: The impact on knowledge creation and dissemination of knowledge generated in the project depends on a high level of activity in dissemination to* *the proper groups.* | * *Number of journal publications acknowledging the project:* **<***input here***>**
* *Number of conference papers and presentations*: **<***input here***>**
 |
| Exploitation plans *(Please describe the exploitation plans related to this Case Study, e.g., summarize the potential stakeholders (public, private, international, etc.) and relate them with the exploitation possibilities)* |
| <*input here*> |

**A.1.3 Case Study**

*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 selection of the use stories shall be representative reflecting both of the research challenge and complexity, and of the possible solutions offered by the Open Data Platform****.* ***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 OpenProject tool[[1]](#footnote-1).*

|  |
| --- |
| ***User Stories (****Please describe use stories, selecting those only related to the Open data platform technology, describe who (actor) wants to do what, need what services/functions and handle what information objects (data, metadata, signals etc., indicate related community policies and constraints, e.g. on data publication, access, preservations, etc.)* |
| * Cryo-EM use case
	+ Videos of between 10-50 frames, each of size 4000x4000 pixeles, are taken at the Electron Microscope
	+ Videos are corrected for within and between frames re.arrangements, producing an “average corrected image”
	+ Subimages of macromolecular complexes are extracted from the average corrected images (or from the frames directly)
	+ Subimages are used for the 3D reconstruction
	+ Transfer the cryo EM videos to the Instruct Image Processing Center in Madrid from several microscopes in Europe
 |

|  |  |
| --- | --- |
| **Information approved by** | <*input here*> |

A.2 Information Viewpoint

*Information viewpoint concerns data object model and data lifecycle in the system. This section of questionnaire should provide the information on the data content, data formats and data lifecycles used in the community without specifying particular technologies and platforms used for data management. Information in this section needs inputs and approvals from data managers of the user community.*

**A.2.1 Data**

|  |
| --- |
| ***Current status*** |
| **Data Object types** (*Please list data object types in current system,* *e.g., level 1 data, level 2 data, raw data, aggregated data, simulation data, etc. and give definition/description of them*) |
| * Raw NMR data (time signals)
* Processed NMR data
* 3D structures (=x,y,z coordinates)
* electron density maps of biomolecules
 |
| **Data size** *(typical size of single file or object)* | * Raw NMR data – 1-50MB per sample
* Processed NMR data – several 100MB
* Analysed NMR data – several GBs
 |
| **Data collection size** *(estimate of total size of data collection in community)* | * Cryo-EM – 2.5TB per day
 |
| **Data format***(e.g. XML, CSV)* | PDB (Protein Data Bank), Text files |
| Data Identifiers *(how is the data objects/files identified)* | **<***input here***>** |
| Standards in use (e.g. FITS, DICOM) | <*input here*> |
| Data locations (&contacts) | <*input here*> |
| Data management plan *(How long should the data be preserved? When can it be made public?)* | * The sample is often the most precious part and is specific to one project and group
* As such data are only made public upon publication
* The only data sharing aspects may-be transferring/sharing data between experimental sites and the user lab(s)

Although the processed / analysed data in structural biology are typically deposited in public databases (e.g. www.pdbe.org, www.bmrb.wisc.edu ), the raw data are in most case only stored locally (often without clear policies or metadata). We foresee a need to long term preservation of raw experimental data. Next to experimental data, there is also a need for open repositories for modelling data (i.e. results of simulations rather than experiments). This can be within institutional repositories (often not yet present), EUDAT or related initiatives. Sharing / preserving data (and making them citable) in the context of a federated data cloud under EGI is a scenario that will need to be investigated.*(from EGI Indigo Data Cloud deliverable D2.1)* |
| **Privacy policy** *(Who can access the data?)* | * Users usually visit the facilities to collect data, and take the data back with them
* User is the owner of the data (INSTRUCT data policy)
* Implementation of data policy for H2020 ongoing (archiving, sharing, metadata,...)
 |
| Other aspects | **<***input here***>** |
| *Future Requirements* |
| **<***input here***>** |

**A.2.2 Metadata**

|  |
| --- |
| ***Current Status*** |
| **Metadata object types** (*Please list metadata object types in current system,* *e.g, metadata for level1 data, metadata for processing data, etc. and give definition/description of them*) |
| <*input here*> |
| Metadata Identifiers | **<***input here***>** |
| Metadata size | <*input here*> |
| Metadata format | **<***input here***>** |
| Standards in use | <*input here*> |
| Metadata generation | **<***input here***>** |
| Metadata locations (&contacts) | **<***input here***>** |
| Other aspects | **<***input here***>** |
| *Future Requirements* |
| **<***input here***>** |

**A.2.3 Data Lifecycle**

|  |
| --- |
| *Current Status* |
| Data Lifecycle (*Please describe the dataflow in current system, indicate explicitly what data object change from which state to which state after what functions/action applied to the data object. E.g., level 1 data become level 2 data after quality checking. Use figure wherever possible.*) |
| **<***input here***>** |
| *Future Requirements* |
| **<***input here***>** |

|  |  |
| --- | --- |
| **Information approved by** | **<***input here***>** |

A.3 TECHNOLOGY Viewpoint

*Technology viewpoint concerns how the data specified in information viewpoint is managed currently in the community. Questionnaire should provide information what technologies are used to store, transfer, access, process and secure the community data sets.*

**A.3.1 General aspects**

|  |
| --- |
| *Current status* |
| System Architecture (*please describe how the functionalities are distributed onto current physical devices, use figure if possible*) |
| MoBRAIN portals are use largely HADDOCK software developed within WeMNR portal. The HADDOCK portal effectively implements a complex workflow in which user data are first validated and processed before HADDOCK computations are launched. Each HADDOCK run correspond to a complex workflow, orchestrated by a master python script that manages the workflow, generates jobs for submission to local queues (e.g. via torque) or grid resources, and monitors the results.  |
| Data management (Please describe how you access and manage your data sets) |
| **Community data access protocols** *(e.g. POSIX, GridFTP, WebDAV)* | FTP |
| **Data management technology** *(Please describe what is the data management system in your community, e.g. LFC, iRODS, etc.)* | <*input here*> |
| **Data access control** *(e.g. POSIX filesystem rights, ACL)*  | <*input here*> |
| **Public data access protocol** *(How should the data be accessed by public users? e.g. HTTP)* | <*input here*> |
| **Public authentication mechanism** *(e.g. anonymous access, track who downloaded file based on X.509 certs)* | * WeNMR SSO - <https://www.wenmr.eu/wenmr/wenmr-sso-module>
* Supports eduGAIN and social login
 |
| Computing capacities *(Please describe the type and capacities of current physical devices used for your data processing)* |
| CPU | <*input here*> |
| GPU | <*input here*> |
| RAM  | <*input here*> |
| Storage *e.g., HDD, tapes* | <*input here*> |
| Network | <*input here*> |
| e-Infrastructure, *e.g., Clusters, Grid, Cloud, Supercomputing resources* | <*input here*> |
| Client, *e.g., workstation, desktop, laptop, Mobile device, etc.* | <*input here*> |
| *Other aspects* | <*input here*> |
| *Future requirements* |
| * Estimated storage space: 100TB
* Estimated CPU requirements: 3000 CPU/year
 |

###

**A.3.2 Non-functional requirements**

*This subsection should provide some information about the non-functional requirements related to data management of the data in the community and in case when the data is made open to the public.*

|  |  |  |
| --- | --- | --- |
| Performance Requirements | Requirement Levels | Description (*please describe performance requirements for the required system*) |
| High | Middle | Normal |
| Availability | <Y/N> | <Y/N> | <Y/N> | <input here> |
| Accessibility | <Y/N> | <Y/N> | <Y/N> | <input here> |
| Throughput | <Y/N> | <Y/N> | <Y/N> | <input here> |
| Response time | <Y/N> | <Y/N> | <Y/N> | <input here> |
| Security | <Y/N> | <Y/N> | <Y/N> | <input here> |
| Utility | <Y/N> | <Y/N> | <Y/N> | <input here> |
| Reliability | <Y/N> | <Y/N> | <Y/N> | <input here> |
| Scalability | <Y/N> | <Y/N> | <Y/N> | <input here> |
| Efficiency | <Y/N> | <Y/N> | <Y/N> | <input here> |
| Disaster recovery | <Y/N> | <Y/N> | <Y/N> | <input here> |
| ***Others performance requirements*** |
| <*input here*> | <Y/N> | <Y/N> | <Y/N> | <*input here*> |
|  |  |  |  |  |
|  |  |  |  |  |

**A.3.3 Software and applications in use**

|  |  |
| --- | --- |
| Software/ applications/services  | * *Describe the software/applications/services name, version:* **<***input here***>**
* *Describe the software licensing:* **<***input here***>**
* *Describe the configuration:* **<***input here***>**
* *Describe the dependencies needed to run the application, indicating origin and requirements:* **<***input here***>**
 |
| Operating system | <*input here*> |
| Runtime libraries/APIs *(e.g., Java, C++, Python, etc.)*  | <*input here*> |
| Typical processing time | <*input here*> |

**A.3.4 e-Infrastructure in use**

|  |
| --- |
| **e-Infrastructure resources being used or planned to be used**. *Please indicate from the point of view of the research community if the current solution is already using an e-Infrastructure (like GEANT, EGI, PRACE, EUDAT, a Cloud provider, etc.) and if so what middleware is used. If relevant, detail which centres support it and what level of resources are used (in terms of million-hours of CPU, Terabytes of storage, network bandwidth, etc.).* |
| <*input here*> |

**A.3.5 Requirements for EGI Testbed Establishments**

|  |
| --- |
| *Does the case include preferences on specific tools and technologies to use? For example: grid access to HTC clusters with gLite; Cloud access to OpenStack sites; Access to clusters via standard interfaces; Access to image analysis tools via Web portal* |
| <*input here*> |
| *Does the user have preferences on specific resource providers? (e.g. in certain countries, regions or sites)*  |
| <*input here*> |
| *Approximately how much compute and storage capacity and for how long time is needed? (may be irrelevant if the activity is for example assessment of an EGI technology)* |
| <*input here*> |
| *Does the user (or those he/she represents) have access to a Certification Authority? (to obtain an EGI certificate)* |
| <*input here*> |
| *Does the user need access to an existing allocation (🡪 join existing VO), or does he/she needs a new allocation? (🡪 create a new VO)* |
| Yes, enmr.eu VO  |
| *Does the user (or those he/she represent) have the resources, time and skills to manage an EGI VO?*  |
| <*input here*> |
| *Which NGIs are interested in supporting this case? (Question to the NGIs)* |
| <*input here*> |

|  |  |
| --- | --- |
| **Information approved by** | <*input here*> |

1. <https://www.openproject.org/> [↑](#footnote-ref-1)