

Requirements Collection

for Open Data Cloud Platform

BBMRI

|  |  |
| --- | --- |
| **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** | Lukasz Dutka |

****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 |
| Template is reviewed by Technology provider | PENDING | Lukasz Dutka | 13 Jul 2015 |
| Requirement extraction from available materials from the community | GATHERING | Bartosz Kryza | 27 Jul 2015 |
| Review of the information collected | REVIEWING | Yin Chen | 27 Jul 2015 |
| Send to the community for confirming | 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** | Biobanking and BioMedical Research Infrastructure |
| Community Short Name if any | BBMRI |
| Community Website | <http://bbmri-eric.eu>  |
| **Community Description**  | 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. |
| **Community Objectives** | * Increase biobank interoperability and data discovery in BBMRI community by providing a secure and standard way to share biobank high-throughput data,
* Provide biobanking community with a federated infrastructure for big data storage and intensive data analysis,
* Facilitate the efficient use of bio-resources by supporting visibility and sharing, while also respecting the protection level required by owners of the data and samples,
* Facilitate the efficient use of economic resources in BBMRI by providing a common infrastructure for storage and processing of big data.
 |
| **Main Contact Institutions** | <*input here*> |
| **Main Contact**(*name and email*) | <*input here*> |
| 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 Platform***) |
| - Assuring confidentiality of donors- Following a user centered approach and providing a maximum support for researchers- Flexibility in terms of biobank content and schema handling- Efficiency in terms of query processing- Low effort for biobanks to participate in federation |
| **Objectives** (*Please describe your objectives to be achieved through collaboration with* ***Open Data Platform****)* |
| * Data from multiple biobanks must be pooled
 |
| 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.)* |
| *(Based on BBMRI deliverable D5.4 Requirements for a general information management system.)*Within an IT-infrastructure for federated biobanks authorized researchers should have the possibility to search and obtain required material and data from all participating biobanks, necessary e.g. to perform biomedical studies. The following section introduces two user scenarios. Based on these scenarios use cases for BBMRI have been derived. Those use cases are presented and discussed in detail in section 3.**User scenario 1: Search for biobanks**1. Peter logs in to the BBMRI portal and goes to the search function.2. He then first needs to choose one of the ontologies that are included in UMLS Metathesaurus:a. Select ontology* ICD-10

3. He makes the following choices of input (by query statements or predefined scrollbars) for the set"Attributes for biobanks":a. Type of diagnosis | Using ICD-10* C50. //Note: non-specific code not used for diagnosis

b. Type of sample* tissue

c. Method of preservation* All

d. Data needed* Diagnosis information

e. Number of samples* Collections with at least 100 samples

4. The client software sends the query for the requested information to the closest BBMRI hub, which is the National Biobank Hub of the Netherlands (NBHNL).5. NBHNL will now:a. Perform a search on the requested input within its own database, which contain "Attributesfor biobanks" for all BBMRI associated biobanks in the Netherlands. UMLS is used to findmappings to samples that have breast cancer as an associated diagnosis defined in analternative ontology.b. Make a request to all other BBMRI hubs to perform a similar search.6. NBHNL puts together the results from all the BBMRI hubs and sends it back to the client software atLeiden University Medical Center.**User scenario 2: Search for samples and/or subjects**8. Starting with the output in Table 1, Peter could now go on and look more specifically at availableattributes for subjects by resolving the collections into subjects.9. Peter makes the following choice:a. Continue* Resolve Biobanks/Collections to Subjects

10. Since he wants to limit the meta analysis to subjects between 30-40, and check that necessaryattributes are available he makes the following choices\* of input for the set "Attributes for subjectsand samples":a. Subject Sex* Available AND Female

b. Subject TNM Staging* Available

c. Subject Age at diagnosis* Available AND 30-40

d. Subject BMI* Available

*\*NOTE: At this stage it must be considered if it should be possible to get source data for certain**attributes (as for "Sex" and "Age at diagnosis") or if only existence can be checked. Privacy protection**could be guaranteed by k-anonymity.*11. By using the result set in Table 1the client software sends the new query to the NBHNL.12. NBHNL will now:a. Perform a search on the requested input on those biobanks that have been identified previously. This means that access have to be granted to individual biobanks within the network.b. Make a request to all other BBMRI hubs to perform a similar search.13. NBHNL puts together the results from all the BBMRI hubs and sends it back to the client software atLeiden University Medical Center. |

|  |  |
| --- | --- |
| **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*) |
| <*input here*> |
| **Data size** *(typical size of single file or object)* | <*input here*> |
| **Data collection size** *(estimate of total size of data collection in community)* | <*input here*> |
| **Data format***(e.g. XML, CSV)* | <*input here*> |
| Data Identifiers *(how is the data objects/files identified)* | DOI (Data Object Identifier), LSID (Life Science Identifier) |
| 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?)* | **<***input here***>** |
| **Privacy policy** *(Who can access the data?)* | <*input here*> |
| 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 | ICD-9, ICD-10, SNOMED CT, UMLS |
| 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*) |
| *(Based on BBMRI deliverable D5.6 Final report)*Within the proposed architecture, a regional hub (autonomous server or biobank) uses a meta structure to provide data sharing. All regional hubs (and participating biobanks that act as hub) are connected via a peer-to-peer structure and communicate with each other via standardized and shared interfaces. Participating European biobanks (without hub function) are connected with its specific regional hub via hub and spoke-structure.  |
| Data management (Please describe how you access and manage your data sets) |
| **Community data access protocols** *(e.g. POSIX, GridFTP, WebDAV)* | <*input here*> |
| **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)* | <*input here*> |
| 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* |
| **<***input here***>** |

###

**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)* |
| <*input here*> |
| *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)