

Requirements Collection

for Open Data Cloud Platform

agINFRA

|  |  |
| --- | --- |
| **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](mailto: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 Cloud 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 |
| Template questions are reviewed by technology provider | PENDING | Lukasz Dutka | 13 Jul 2015 |
| Information collection based on available material | GATHERING | Bartosz Kryza | 20 Jul 2015 |
| Information reviewed by internal team | 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** | EGI Federated Cloud services for the agri-food research |
| Community Short Name if any | agINFRA |
| Community Website | http://www.aginfra.eu |
| **Community Description** | The [agINFRA](http://aginfra.eu/) project, supported by the Agriculture Information Management Standards of the Food and Agriculture Organization of the United Nations ([AIMS FAO](http://aims.fao.org/)) and the [CIARD](http://www.ciard.net/) 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](http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf) mandate and share their metadata with their thematic aggregator in order to publish them in [OpenAire](http://openaire.eu/). *(from www.aginfra.eu)* |
| **Community Objectives** | agINFRA aims to function as the thematic aggregator of the agri-food research domain and act as the main research community for OpenAire. |
| **Main Contact Institutions** | Agro-Know, FAO |
| **Main Contact**  (*name and email*) | <*Effie Tsiflidou effie@agroknow.gr*> |
| Prior requirement capture activities and ideally a summary and references to their outcome | agINFRA D2.2 Revised stakeholders needs deliverable [*http://www.aginfra.eu/project/images/DELIVERABLES/aginfra\_d2.2\_revised-review-of-stakeholder-needs\_final\_20131025.pdf*](http://www.aginfra.eu/project/images/DELIVERABLES/aginfra_d2.2_revised-review-of-stakeholder-needs_final_20131025.pdf)  *agINFRA D5.5 Report on agricultural data sources/repositories integration* |
| 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***) | | |
| * High volume storage * Impossible to use centralized storage * Large, live, constantly updated data streams * Handling of heterogeneous data | | |
| **Objectives** (*Please describe your objectives to be achieved through collaboration with* ***Open Data Platform****)* | | |
| * Raw data resources with agricultural data must be publicly available, using a unified search and discovery platform * Making such resources more broadly discoverable by humans and machines by registering them in shared public directories and providing all the technical information that allows applications to process those data * Reach out to entrepreneurs who can put their data to work in new services * Invite commercial entities into the conversation around the future of data | | |
| Expectations *(please describe your expectations for the new technology to be provided by the* ***Open Data Platform****)* | | |
|  | | |
| 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: <10000 monthly>* * *Number of sites provide the services:* <*20*> |
| *Usability* | *Simplifying deployment of the web portals in cloud resources* | * *Number of downloads:* **<***we don’t keep analytics for downloads***>** |
| *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:* **<***4***>** * *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.)* |
| Use cases taken from agINFRA public deliverable D1.3.3 agINFRA Scientific Vision: Part A   * Data provider who needs to host and store a small scale CMS In this case, data provider requests from the system to set up his own CMS instance in order to cover the needs for a small scale CMS E.g. Open Educational Resources (http://www.oercommons.org/), which provides access to hundreds of course-related materials and collections in several themes * Data provider, who needs to host and store a large scale hosting & replication CMS In this case, data provider requests from the system to allocate space or to set up accounts in a large scale CMS E.g. Consiglio per la Ricerca e la Sperimentazione in Agricoltura - CRA (http://sito.entecra.it/portale/index2.php), which includes thousands of data sources in several research fields in agriculture and related domains * Data provider, who needs to host CMS at own or external / commercial infrastructure In this case, content provider is interested to expose (meta)data to e-infrastructure, E.g. Turkish Agricultural Learning Objects Repository - TrAgLOR (http://traglor.cu.edu.tr/), which serves as an organized collections of learning objects, stored on servers and delivered through networks. |
|  |

|  |  |
| --- | --- |
| **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*) | |
| * Germplasm data | |
| **Data size** *(typical size of single file or object)* | ~ 10KB |
| **Data collection size** *(estimate of total size of data collection in community)* | ~ 1PB |
| **Data format**  *(e.g. XML, CSV)* | XML |
| Data Identifiers *(how is the data objects/files identified)* |  |
| Standards in use (e.g. FITS, DICOM) | MCPD (for Germplasm data) |
| Data locations (&contacts) | <*cloud infrastructure*> |
| Data management plan *(How long should the data be preserved? When can it be made public?)* | * agINFRA collects data free of access to make them publicly available * agINFRA should ensure long-term preservation |
| **Privacy policy** *(Who can access the data?)* | * publicly available, free of access |
| 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*) | |
| * AGRIS Bibliographic information: **metadata** for publications (scientific articles, thesis, dissertations, journals) * GLN metadata for educational resources. * VocBench instances * VEST Registry * CIARD RING | |
| Metadata Identifiers | ARN |
| Metadata size | ~ 10KB |
| Metadata format | RDF, OWL, XML |
| Standards in use | RDF, OWL, SKOS, OAI-PMH |
| Metadata generation | **<***custom java code based on xml transformations***>** |
| Metadata locations (&contacts) | URI of the AGRIS resources |
| Other aspects | triple store with RDF files in order to preserve linked open data |
| *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.*) |
| Data acquisition level (including manual sent raw XML files or harvesting via protocols like OAI-PMH)  Metadata records evaluation and mappings  Data transformation  Data identification – deduplication  Data triplification (XML to RDF)  Upload RDFs to allegro-graph triple store  Data indexing  Data publishing to AGRIS portal and also provide an FTP with XML records and RDFs  \*Data curation |
| *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*) | |
| In the context of the agINFRA project, there are a number of data providers providing access to different data types, such as educational, bibliographic, germplasm, statistical, soil maps, cultural and other. The aggregation of metadata from these data sources, which use different metadata schemas in order to meet the specific requirements of each data type, would traditionally be carried out by individually transforming and then harvesting each data source. This approach would be most appropriate for serving the data integration as well as other services deployed by the agINFRA project. A more state-of-the-art methodology should apply the current advances in the context of the Semantic Web, including the publication of all available data as linked and open data. The first step in this process would be the development of a metadata model for each resource type, which would accommodate the most common and / or essential elements of the metadata schemas used in agINFRA by the data providers. *(from agINFRA D5.4 public deliverable)*  agINFRA infrastructure pays special attention in topics like the efficient metadata management (checking for mappings and transformation of the targeted metadata schemas to a common schema), storage issues for hosting data components and scaling up the handled metadata aggregations and their versions, computing issues in terms of time and resources that are needed for harvesting and often recurring for the coverage similar workflows that are needed (for validation, transformation, harvesting, auto-tagging and indexing). *(from agINFRA D1.3.3 public deliverable)* | |
| Data management (Please describe how you access and manage your data sets) | |
| **Community data access protocols** *(e.g. POSIX, GridFTP, WebDAV)* | web interface & FTP |
| **Data management technology** *(Please describe what is the data management system in your community, e.g. LFC, iRODS, etc.)* | Custom |
| **Data access control** *(e.g. POSIX filesystem rights, ACL)* | POSIX |
| **Public data access protocol** *(How should the data be accessed by public users? e.g. HTTP)* | HTTP |
| **Public authentication mechanism** *(e.g. anonymous access, track who downloaded file based on X.509 certs)* | *anonymous access* |
| Computing capacities *(Please describe the type and capacities of current physical devices used for your data processing)* | |
| CPU | 3500 CPU’s |
| GPU | No |
| RAM | 4GB |
| Storage *e.g., HDD, tapes* | 30GB |
| Network |  |
| e-Infrastructure, *e.g., Clusters, Grid, Cloud, Supercomputing resources* | Cloud |
| Client, *e.g., workstation, desktop, laptop, Mobile device, etc.* | Desktop, laptop, mobile device |
| *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 | N | <input here> |
| Accessibility | Y | N | N | <input here> |
| Throughput | Y | N | N | <input here> |
| Response time | Y | N | N | <input here> |
| Security | Y | N | N | <input here> |
| Utility | Y | N | N | <input here> |
| Reliability | Y | N | N | <input here> |
| Scalability | Y | N | N | <input here> |
| Efficiency | Y | N | N | <input here> |
| Disaster recovery | Y | N | 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:* **<apache tomcat, solr, custom java code>** * *Describe the software licensing:* **<open source>** * *Describe the configuration:* **<***our web app is based on java war application and run on tomcat 6***>** * *Describe the dependencies needed to run the application, indicating origin and requirements:* **<***cloud infrastructure, open jdk 6, apache tomcat 6***>** |
| Operating system | <*centos 5 linux*> |
| Runtime libraries/APIs *(e.g., Java, C++, Python, etc.)* | <*java, sax parser, solr 1.4*> |
| Typical processing time | <*15h*> |

**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.).* |
| EGI, GEANT though GRNET cloud (oceans and Vima) |

**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* |
| <*cloud infrastructure like virtual machine instances*> |
| *Does the user have preferences on specific resource providers? (e.g. in certain countries, regions or sites)* |
| <*no*> |
| *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)* |
| <*2.2 GHz, long term preservation, 100 GB*> |
| *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)* |
| <*no*> |
| *Does the user (or those he/she represent) have the resources, time and skills to manage an EGI VO?* |
| <No> |
| *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)