

Requirements Collection

for Open Data Cloud Platform

CANFAR

|  |  |
| --- | --- |
| **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 | 4 Jul 2015 |
| Distribute the template to community | GATHERING | Yin Chen, Bartosz Kryza | 17 Jul 2015 |
| Interview selected communities | GATHERING | Lukasz Dutka, Bartosz Kryza, Yin Chen |  |
| Requirements reviewed by internal team | REVIEWING | Tiziana Ferrari, Lukasz Dutka, (others) |  |
| Send to the community for confirming | CONFIRMING | Bartosz Kryza, Yin Chen |  |
| Get approvals from the community | ACCEPTED |  |  |
| Complete information collection | COMPLETE | Bartosz Kryza, Yin Chen |  |

* **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** | Canadian Advanced Network for Astronomical Research |
| Community Short Name if any | CANFAR |
| Community Website | <http://www.canfar.phys.uvic.ca/canfar/> |
| **Community Description** | The Canadian 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. *(From http://www.canfar.net/about)* |
| **Community Objectives** | 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. |
| **Main Contact Institutions** | Instituto Nazionale di Astrofisica (INAF), Via G.B. Tiepolo, 11 I-34143 Trieste, Italy - Tel. +39 040 3199 111 - [infoats@oats.inaf.it](mailto:infoats@oats.inaf.it) |
| **Main Contact**  (*name and email*) | Giuliano Taffoni <taffoni@oats.inaf.it> |
| Prior requirement capture activities and ideally a summary and references to their outcome | N/A |
| 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 main problem related to CANFAR case study is that European A&A community has only storage infrastructure, without computing, which is available in the Canadian A&A cloud. The typical observation files are very large, and thus very expensive to transfer to computational sites. After the data is made public (typically after 1 year) it should be replicated between European and Canadian cloud storage. | | |
| **Objectives** (*Please describe your objectives to be achieved through collaboration with* ***Open Data Cloud****)* | | |
| The main objectives of this community with respect to Open Data Cloud include:   * Establish close collaboration between European and Canadian astronomy and astrophysics (A&A) communities, * Enable sharing large volumes of astronomical observation and simulation data according to agreed policies (e.g. after 1 year of publication data should be public) * Enable replication of data between Canadian and European Cloud storage infrastructures. | | |
| Expectations *(please describe your expectations for the new technology to be provided by the* ***Open Data Cloud****)* | | |
| <*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.)* |
| **UC1: User data is made public automatically after 1 year.**  **Actors:**   * Principal Investigator who created the original data.   **Action:**   * Access to the data is automatically enabled after 1 year from creation by the Open Data Platform * Data is replicated between CANFAR and EGI infrastructures * Data is available through EGI Open Data Platform portal   **Current solutions:**   * Data transfers are initiated manually   **Problems to be solved:**  **UC2: User wants to find publicly available data set.**  **Actors:**   * Community user interested in accessing particular observation data set.   **Action:**   * User enters in the community portal query specifying selected FITS metadata key/value pairs. Matching data sets are located, and filtered based on the privacy ACL’s set for data (all public data sets matching the query will be returned to any user.)   **Problems to be solved:**   * Enable automatic ACL modification based on time of data creation * Enable public access over specified protocols (e.g. HTTP, FTP) of the public files |
|  |

|  |  |
| --- | --- |
| **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*) | |
| * Astronomical and astrophysical observation raw data (FITS format, includes ASCII header and binary CCD data) * Astronomical and astrophysical observation pre-processed data (e.g. optimized volume) * Astronomical and astrophysical simulation data | |
| **Data size** *(typical size of single file or object)* | ~1TB (one night observation) |
| **Data collection size** *(estimate of total size of data collection in community)* | 1PB |
| **Data format**  *(e.g. XML, CSV)* | FITS (Flexible Image Transport System) |
| Data Identifiers *(how is the data objects/files identified)* | Metadata in the files is located in the headers of FITS files, and also indexed in external SQL database for lookup. |
| Standards in use (e.g. FITS, DICOM) | FITS |
| Data locations (&contacts) | Italian sites and Canadian sites |
| Data management plan *(How long should the data be preserved? When can it be made public?)* | Data is typically owned by Principal Investigator for 1 year, after which it should be made public. The PI can also process the data, pre-process it to reduce its volume. |
| **Privacy policy** *(Who can access the data?)* | For 1 year after creation the policy is defined by the Principal Investigator, i.e. she can decide who can access the data. After 1 year the data should be publicly available. |
| 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*) | |
| Metadata for FITS files is stored in an ASCII header of each file in a simple list of key/value pairs with optional comments. | |
| Metadata Identifiers | Metadata is stored in key/value pairs, metadata identifiers are simple abbreviated strings, e.g. ORIGIN, LPKTTIME, NAXIS, etc. |
| Metadata size | Small in comparison to actual data, typically up to a 100 key/value pairs per file |
| Metadata format | ASCII text header in the beginning of each FITS file, with key/value pairs with optional comments. |
| Standards in use | FITS |
| Metadata generation | Mostly automatic |
| Metadata locations (&contacts) | Metadata is located in the header of each FITS file, as well as indexed in relational databases for data discovery. |
| 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*) | |
| CANFAR (Canadian Advanced Network for Astronomical Research) is composed of:   * Canadian National Research Network (CANARIE) * Cloud processing and storage (Cloud Canada) * Canadian Astronomy Data Centre (CADC)   Together they provide a platform for distribution, processing and storage of astronomical and astrophysical data sets. The cloud infrastructure is based on OpenStack technology. Main service provided to the users include:   * VOSpace – Virtual Observatory user storage, * VMOD – Virtual Machines on Demand, * GMS – Batch processing and group management,   All services are based on RESTful protocols maintained by CADC.  Macintosh HD:Users:morgan:Desktop:2015-04_STScI.pptx.pdf  VOSpace provides a web based user interface for finding datasets based on FITS metadata queries. The metadata from FITS file headers is indexed in a relational database. Users information is stored in an LDAP catalogue. | |
| Data management (Please describe how you access and manage your data sets) | |
| **Community data access protocols** *(e.g. POSIX, GridFTP, WebDAV)* | * REST or SOAP for data management control * HTTP, FTP for data transfers |
| **Data management technology** *(Please describe what is the data management system in your community, e.g. LFC, iRODS, etc.)* | CANFAR data management system is based on VOSpace which is an implementation of Virtual Observatory Specification Draft (<http://www.ivoa.net/documents/VOSpace/20150601/VOSpace.pdf>).  Data management control is available through a RESTful interface. |
| **Data access control** *(e.g. POSIX filesystem rights, ACL)* | GMS service provides the role of Policy Information Point during authorization requests, returning information about users groups, capabilities and capacities. VOSpace permissions are similar to POSIX based rights. |
| **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* | * CANFAR infrastructure is based on OpenStack cloud platform |
| 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 |  |  | <*input here*> |
| Accessibility | Y |  |  | The public data should be easily accessible to all users. |
| Throughput | Y |  |  | Data transfers should use all available bandwidth whenever possible. This can be achieved by striping data into blocks and serving them simultaneously from several nodes in the cluster. |
| Response time |  | Y |  | * Response time in terms of metadata queries should be quick in terms of typical user experience * Response time in case of large data set transfer is not critical (data transfers will take several minutes/hours anyway) |
| Security | Y |  |  | Only data which is publicly available should be accessible by non-authorized users. |
| Utility |  |  |  | <*input here*> |
| Reliability | Y |  |  | <*input here*> |
| Scalability | Y |  |  | <*input here*> |
| Efficiency |  |  |  | <*input here*> |
| Disaster recovery |  |  |  | <*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.).* |
| * EGI FedCloud * CANFAR (OpenStack) |

**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* |
| * Automatic provision of public data sets to users (based on predefined policies, e.g. after 1 year since creation) * Due to large size of data sets, data transfer from storage site to computation site can be very expensive. Either computation should be moved closed to data, or if not possible, local mount of the remote storage on the computational nodes should be provided |
| *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)* |
| Current data size is over 1PB in size. |
| *Does the user (or those he/she represents) have access to a Certification Authority? (to obtain an EGI certificate)* |
| This will be resolved as part of EGI FedCloud project. Authentication will be based on X.509 certificates and in the future possibly based on eduGAIN service. |
| *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)