

Requirements Collection

for Open Data Cloud Platform

CTA

|  |  |
| --- | --- |
| **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 |
| Template questions reviewed by technology provider | PENDING | Lukasz Dutka | 13 Jul 2015 |
| Information collection based on available material | GATHERING | Bartosz Kryza | 27 Jul 2015 |
| Requirements 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** | Cherenkov Telescope Array (CTA) Observatory |
| Community Short Name if any | CTA |
| Community Website | http://www.cta-observatory.org |
| **Community Description**  | 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. |
| **Community Objectives** | 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.
 |
| **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***) |
| * CTA will produce unprecedented amount of data (>1000PB)
* CTA data and their scientific products need to be preserved in a dedicated archive whose aim is to provide open access to a wide and diverse scientific community for several years after the end of the CTA operative life (~30years)
* Enable data dissemination to the scientific community through Science Gateway and Single Sign-On solutions for this huge amount of data
 |
| **Objectives** (*Please describe your objectives to be achieved through collaboration with* ***Open Data Platform****)* |
| * Enable open access to the large volumes of CTA data
 |
| 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.)* |
| 1. Guest Observer (GO, i.e. Principal Investigator) will need to browse proprietary data related to a private proposal and after few interactions with a web-based system will retrieve the whole set of data and prepare several higher level outputs to be published in the VO. → The underlying archive will be distributed and data will be sent to the nearest storage node according to the GO location in order to minimize latency of further computation and analysis on data. Thus a data retrieval system will grant the peer-to-peer or torrent download on local GO machine.
2. PIPELINE user (PU) will need to access RAW and intermediate data from the archive in order to perform reduction tasks and will store outputs in the archive as higher data levels. → the same distributed archive will handle the authentication of a privileged user, which will run pipelines scripts job on archive/computing cores in the place where data are stored and will prompt the output to the PU user. A scheduler-Job manager will take care of redirect and queue jobs and dag on the proper archive IT- resource.
3. Data Ingestion of each RAW data flow, as well as pipeline intermediate/final data products as well as simulation MC data or just information on P.I. proposals → a dedicated software tool will take care to handle and organize the archive data and meta-data descriptors related to data in a distributed environment and the associated database will be distributed too.
4. Scheduler and TAC users will only need to validate the GO proposals and thus grade and rank them in order to best schedule and queue for observation → this action will mailnly performed through a dedicate web portal access included in the User access section of the CTA Archive.
5. Archive User will take care of best balance archive repositories and database workload in order to optimize all the archive data flow and solve issues → this action can be orchestrated at the Archive, DB and OS Management System, using dedicated tools.
 |

|  |  |
| --- | --- |
| **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 data – raw components of EVT and CAL data stream (EVT0 and CAL0)
* Calibration data – CAL1 (single telescope calibration data) and CAL2 (array calibration data)
* Engineering data
* MC data – Monte Carlo simulation data
* Science data – accessible in a controlled way by Guest Observers
* High Level Multi-Frequency Data
 |
| **Data size** *(typical size of single file or object)* | <*input here*> |
| **Data collection size** *(estimate of total size of data collection in community)* | >1000PB (target size) |
| **Data format***(e.g. XML, CSV)* | FITS, ROOT, RAW, JSON, XML, BSON |
| Data Identifiers *(how is the data objects/files identified)* | The complexity of astronomical data is difficult to be addressed by a an Object Identifier but it is possible to tune-up the searchable keywords for VO research in order to create an ad hoc Astronomical Object Identifier.  |
| Standards in use (e.g. FITS, DICOM) | FITS |
| Data locations (&contacts) | <*input here*> |
| Data management plan *(How long should the data be preserved? When can it be made public?)* | 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 to 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.  |
| **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 | <*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*) |
| The CTA Archive System (CTAAS) architecture is composed of 2 main subsystems:* An on-site archive that represents the front-end CTAAS interface to CTA arrays. It is composed of appropriate repository units that will temporary store raw data files (DL0), the engineering data (TECH) and a set of pre-defined LUT/IRFs (and/or a sub- sample of MC files for their on-site production). The archive will also host the data products expected by the on-site and real-time analysis, including preliminary science products. It will implement management system including databases and access tools for archive browsing by CTA internal users, CTA operators and by the Data Processing Pipelines. It can be considered as one of the node since the whole CTAAS could be federated in a cloud.
* An off-site archive that will permanently store and make available all the data and observatory products for the scientific community, through dedicated services and databases

 |
| 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 | ~2000 cores / 10-20 GPUs servers  |
| GPU | <*input here*> |
| RAM  | 4-8GB per thread/core  |
| Storage *e.g., HDD, tapes* | >100PB |
| Network | >10GB/s point to point |
| 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  | *C/C++ Java, Python with different libraries, CFITSIO, GSL, DJANGO framework (python) and Liferay framework (Java)*  |
| Operating system | LINUX (scientific Linux) for servers, No restriction for clients  |
| Runtime libraries/APIs *(e.g., Java, C++, Python, etc.)*  | C/C++, Java & Python  |
| 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.).* |
| Since 2008 a dedicated CTA Computing Grid (CTACG) was created in order to run simulations. The related CTA Virtual Organisation (CTA VO), support 18 grid sites spread in seven countries, with resources of the order of some thousands of available logical CPUs and more than 600 TB of storage. To work in production it is recommended to enhance the virtual organisation and spread it over more countries. To gather different services it is also planned to migrate to a cloud environment.  |

**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, CTA 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)