

Requirements Collection

for Open Data Platform

LoFAR

|  |  |
| --- | --- |
| **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 |
| Community Representative | Michael Wise | ASTRON | wise@astron.nl |
| Community Representative | R.F.Pizzo | ASTRON | pizzo@astron.nl |
| Community Representative | Susana Sanchez-Exposito | CSIC | sse@iaa.es |
| Community Representative | Daniele Lezzi | BSC | daniele.lezzi@bsc.es |
| Community Representative | Jose Sabater Montes | IAA | jsm@iaa.es |

****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 |
| Get approvals of the technical details of the template | PENDING | Lukasz Dutka | 13 Jul 2015 |
| Information filled based on available materials resources | GATHERING | Yin Chen | 27 Jul 2015 |
| Requirements reviewed by internal team | REVIEWING | Bartosz Kryza | 28 Jul 2015 |
| Send to the community for providing missing information and confirming | CONFIRMING | Yin Chen | 28 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** | The Low Frequency Array |
| Community Short Name if any | LoFAR |
| Community Website | <http://www.lofar.org/> |
| **Community Description** | LoFAR will be the first large radio telescope system wherein a huge amount of small sensors are used to achieve its sensitivity instead of a small number of big dishes. For the astronomy application, LOFAR is an aperture synthesis array composed of phased array stations. The antennas in each station form a phased array, producing one or many station beams on the sky. Multi-beaming is a major advantage of the phased array concept. It is not only used to increase observational efficiency, but may be vital for calibration purposes. The phased array stations are combined into an aperture synthesis array. The Remote Stations are distributed over a large area with a maximum baseline of 100 km within the Netherlands and 1500 km within Europe. |
| **Community Objectives** | LOFAR started as a new and innovative effort to force a breakthrough in sensitivity for astronomical observations at radio-frequencies below 250 MHz |
| **Main Contact Institutions** | ASTRON, CSIC, BSC, IAA |
| **Main Contact**  (*name and email*) | Michael Wise (ASTRON, [wise@astron.nl](mailto:wise@astron.nl)),  R.F.Pizzo (ASTRON, [pizzo@astron.nl](mailto:pizzo@astron.nl) ),  Susana Sanchez-Exposito (CSIC, [sse@iaa.es](mailto:sse@iaa.es))  Daniele Lezzi (BSC, [daniele.lezzi@bsc.es](mailto:daniele.lezzi@bsc.es))  Jose Sabater Montes (IAA, [jsm@iaa.es](mailto:jsm@iaa.es)) |
| Prior requirement capture activities and ideally a summary and references to their outcome | EGI and LoFAR have been collaborating since Oct 2014 to integrate calibration, analysis and modelling pipelines of radio-astronomy data into a cloud infrastructure. It is developed jointly by users of the [www.lofar.org LOFAR] radio-telescope and members of the [AMIGA4GAS](http://amiga.iaa.es/) project. |
| Upload copies of files and provide links to them | <https://wiki.egi.eu/wiki/FedCloudLOFAR> |
| 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***) | | |
| 1. In most cases, ingest jobs by the Radio Observatory need to be monitored closely to verify that all files are ingested and to manually recover the situation after a failure. This causes quite some inconvenience for some users, who have to wait for several days to get their data. 2. Instability of the ingest system can cause long ingest queues and, inevitably, can make CEP2 very full. In extreme cases, the observing schedule needs to be rearranged because there is not enough disk space available on CEP2 to store more data till important ingest jobs are completed and the corresponding data can be removed from the cluster. This obviously limits the observing efficiency. 3. Larger file number/size for staging required 4. Fully exploit processing resources offered by the LOFAR Long-Term Archive | | |
| **Objectives** (*Please describe your objectives to be achieved through collaboration with* ***Open Data Platform****)* | | |
| * Efficient user data retrieval.   1. Optimise the data staging, e.g., using pre-staging technology to move data from tapes to computing facilities to reduce the waiting for staging when a user requests to retrieve data   2. Allowing user to process large amount of data and retrieve results only to avoid downloading the data to their local computer. * Elastic disk storage space to allow data ingest jobs to be smoothly executed when handling bust computation | | |
| 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*) | | |
| Allow users to efficiently access LoFAR Data | | |
| *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.)* |
| LOFAR requests a future system to efficiently support large volumes of data access and burst data access, in particularly, to support the following 2 scenarios:   * User wants to retrieve large volumes of data from LTA. He finds the desired datasets by using the searching facilities provided by LoFAR data portal. An optimisation mechanism is installed which accelerates data staging process. User also starts data processing/analysing service/application running at LTA HPC/Cloud which are near the datasets. The dataset is injected into the processing/analysing service/application, and produced the results. User examines the results using a visualisation service, and downloads the results on his local PC. * LTA encounters a burst access from users, and existing disk space is too small to handle the requests. Since LTA is federated with EGI Cloud, additional resources are immediately assigned to LOFAR LTA to handle the burst access. After that, the additional resources from EGI FedCloud are released. |

|  |  |
| --- | --- |
| **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*) | |
| * Imaging data * Pulsar data | |
| **Data size** *(typical size of single file or object)* | Observational data at rates up to 60 Gbps (650 TB per day), once processed, the amount of data to be kept for a longer time |
| **Data collection size** *(estimate of total size of data collection in community)* | * Exceeded 19 PB of data in the Long-Term Archive (LTA) * Current growth: 3PB per year |
| **Data format**  *(e.g. XML, CSV)* | Datacubes (3D data): two Fourier spatial coordinate axes plus a spectral axis. A datacube can reach several TB. LOFAR telescope allows up to 488 subbands, which can reach several GBs. Each subband is processed independently. |
| Data Identifiers *(how is the data objects/files identified)* | <*input here*> |
| Standards in use (e.g. FITS, DICOM) | <*input here*> |
| Data locations (&contacts) | Currently involves sites in the Netherlands and Germany |
| Data management plan *(How long should the data be preserved? When can it be made public?)* | LoFAR data made public as of March 2nd 2015. |
| **Privacy policy** *(Who can access the data?)* | * Data that has passed the proprietary period becomes public and can be retrieved by anyone. * Currently, data are still mainly  retrieved by project PIs and collaborators |
| 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.*) |
| The LOFAR SYSTEM: Data Flow |
| *Future Requirements* |
| The LOFAR Archive stores data on magnetic tapes. Data cannot be downloaded right away, but has to be copied from tape to disk first. This process is called 'staging’  Current limitations:   * Stage no more than 10 TB at a time and no more than 20,000 files * Staging data from tape to disk might take drives are shared with all users (also non- LOFAR) and requests are queued * Staging space is limited and shared between all LOFAR users – system might temporarily run low on disk space * Data copy remains on disk for 2 weeks * Maintenance and small outages experienced regularly   Require efficient solutions for data retrieval |

|  |  |
| --- | --- |
| **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*) | |
| Architecture of LoFAR Long-Term Archive (LTA) and Web based download server   * Distributed information system created to store and process the large data volumes generated by the LoFAR radio telescope * Currently involves sites in the Netherlands and Germany * Each site involved in the LTA provides storage capacity and optionally processing capabilities. | |
| Data management (Please describe how you access and manage your data sets) | |
| **Community data access protocols** *(e.g. POSIX, GridFTP, WebDAV)* | GridFTP   * Requires grid user certificate * More robust, superior performance * Requires grid client installation |
| **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)* | Web based download server   * ‘LTA enabled’ ASTRON/LOFAR account * Low threshold * Primarily for few files & smaller volumes |
| **Public authentication mechanism** *(e.g. anonymous access, track who downloaded file based on X.509 certs)* | Requires grid user certificate |
| 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 | Network consisting of light-path connections (utilizing 10 GbE technology) that are shared with LOFAR station connections and with the European eVLBI network. |
| e-Infrastructure, *e.g., Clusters, Grid, Cloud, Supercomputing resources* | Grid is in use |
| Client, *e.g., workstation, desktop, laptop, Mobile device, etc.* | Web interface |
| *Other aspects* | Interface to query the LTA database and retrieve data to own compute facilities |
| *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 | Not essential at this moment |
| Accessibility |  |  | Y | Not essential at this moment |
| Throughput |  |  | Y | Not essential at this moment |
| Response time | Y |  |  | Request to reduce staging time for large dataset, support of burst access |
| Security |  |  | Y | Not essential since LOFAR data are open |
| Utility |  | Y |  | LOFAR data shall be used by more users, at the moment main accesses are from PIs. |
| Reliability | Y |  |  | The ingest system is instable which can cause long ingest queues. |
| Scalability | Y |  |  | Request to reduce staging time for large dataset, support of burst access |
| Efficiency | Y |  |  | LOFAR data shall be easily and efficiently accessed |
| Disaster recovery |  |  | Y | Not essential at this moment |
| ***Others performance requirements*** | | | | |
| Flexibility | Y |  |  | The LOFAR pipeline framework is not flexible |
| Decentralisaion | Y |  |  | LOFAT LTAs are decentralised, thus need decentralised solutions |

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

|  |  |
| --- | --- |
| Software/ applications/services | * *Describe the software/applications/services name, version:* Standardized LOFAR pipeline software integration with catalog & user interfaces * *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.).* |
| Grid clusters in use   * SARA * NIKHEF * RUG * FZ-Jülich   Service to LoFAR users   * Standardized pipelines * Integration with catalog & user interfaces * Processing where the data is high complexity & inhomogeneity   Expert users can   * Run custom software * Use native protocols * Build on integration with catalog   + Queries   + Ingest output including data lineage |

**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)* |
| Netherlands, Germany |
| *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)* |
| Yes, requires grid user certificate |
| *Does the user need access to an existing allocation (🡪 join existing VO), or does he/she needs a new allocation? (🡪 create a new VO)* |
| Do not need a new 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)* |
| Amsterdam Sara |

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

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