

Requirements Collection

for Open Data Platform

Human Brain Project

|  |  |
| --- | --- |
| **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 | Sean Hill | Humanbrainproejct.eu | sean.hill@epfl.ch |
| Community Representative | Jeff Muller | Humanbrainproejct.eu | jeffrey.muller@epfl.ch |
| Community Representative | Catherine Zwahlen | Humanbrainproejct.eu | catherine.zwahlen@epfl.ch |

****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 | 14 Jul 2015 |
| Requirements reviewed by internal team | REVIEWING | Yin Chen | 16 Jul 2015 |
| Send to the community for providing missing information and confirming | CONFIRMING | Yin Chen | 16 Jul 2015 |
| Get approvals from the community | ACCEPTED | Catherine Zwahlen,  Jeff Muller | 24 Jul 2015 |
| 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** | Human Brain Project |
| Community Short Name if any | HBP |
| Community Website | [www.humanbrainproject.eu](http://www.humanbrainproject.eu) |
| **Community Description** | The aim of the Human Brain Project (HBP) is to accelerate our understanding of the human brain by integrating global neuroscience knowledge and data into supercomputer-based models and simulations. This will be achieved, in part, by engaging the European and global research communities using six collaborative ICT platforms: Neuroinformatics, Brain Simulation, High Performance Computing, Medical Informatics, High Performance Computing, Neuromorphic Computing and Neurorobotics. |
| **Community Objectives** | In the United States, the **B**rain **R**esearch through **A**dvancing **I**nnovative **N**eurotechnologies (BRAIN) Initiative aims to accelerate the development of new technologies to create large-scale measurements of the structure and function of the brain. The aim is to enable researchers to acquire, analyze and disseminate massive amounts of data about the dynamics nature of the brain from cells to circuits and the whole brain.  For the HBP Neuroinformatics Platform, a key capability is to deliver multi-level brain atlases that enable the analysis and integration of many different types of data into common semantic and spatial coordinate frameworks. Because the data to be integrated is large and widely distributed an infrastructure that enables “in place” visualization and analysis with data services co-located with data storage is requisite. Providing a standard set of services for such large data sets will enhance data sharing and collaboration in neuroscience initiatives around the world. |
| **Main Contact Institutions** | Center for Brain Simulation, Campus Biotech chemin des Mines, 9, CH-1202 Geneva, Switzerland |
| **Main Contact**  (*name and email*) | Sean Hill ([sean.hill@humanbrainproject.eu](mailto:sean.hill@humanbrainproject.eu), +41 21 693 96 78)  Jeff Muller ([jeffrey.muller@epfl.ch](mailto:jeffrey.muller@epfl.ch) )  Catherine Zwahlen ([catherine.zwahlen@epfl.ch](mailto:catherine.zwahlen@epfl.ch) )  Stiebrina Dace ([dace.stiebrina@epfl.ch](mailto:dace.stiebrina@epfl.ch), Secretary ) |
| Prior requirement capture activities and ideally a summary and references to their outcome | EGI and HBP has previous discussions about collaborations in various meetings and in EGI 2015 Lisbon, Towards Open Data Cloud session. |
| Upload copies of files and provide links to them | <https://documents.egi.eu/public/ShowDocument?docid=2468> |
| 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***) | | |
| Open Data Challenges in HBP:   * Not black and white, not just OPEN or CLOSED, need granularity to be explicit about what is open, when and for what purpose, then gradually develop the culture of loosening these restrictions. * Willing to share data but expensive to produce (intellectual capital-experimental design, acquisition cost, time); many possible uses (multiple research questions) for large datasets; currently, reward currencies are intellectual advances, publications and citations; No clear reward or motivation for providing data completely free of any constraint * Willing to share data, need to provide incentives for contributions; establish common data use agreements; adopt common metadata, vocabularies, data formats/services; streamlining deployment of infrastructure to data sources (heterogeneous data access methods/ authentication/authorization); deploying data-type specific services attached to repositories | | |
| **Objectives** (*Please describe your objectives to be achieved through collaboration with* ***Open Data Platform****)* | | |
| **Use Story I – Remote interactive multiresolution visualization of large volumetric datasets**  Large amounts of image stacks or volumetric data are produced daily at brain research sites around the world. This includes human brain imaging data in clinics, connective data in research studies, whole brain imaging with light-sheet microscopy and tissue clearing methods or micro-optical sectioning techniques, two-photon imaging, array tomography, and electron beam microscopy.  A key challenge in make such data available is to make it accessible without moving large amounts of data. Typical dataset sizes can reach in the terabyte range, while a researcher may want to only view or access a small subset of the entire dataset.  ***An active repository*** The ability to easily deploy an active repository that combines large data storage with a set of computational services for accessing and viewing large volume datasets would address a key challenge present in modern neuroscience and across other domains.  **Use Story II – Feature extraction and analysis of large volumetric datasets**  Neurons are essential building blocks of the brain and key to its information processing ability. The three-dimensional shape of a neuron plays a major role in determining its connectivity, integration of synaptic input and cellular firing properties. Thus, characterization of the 3D morphology of neurons is fundamentally important in neuroscience and related applications. Digitization of the morphology of neurons and other tree-shape biological structures (e.g. glial cells, brain vasculatures) has been studied in the last 30 years. Recent big neuroscience initiatives worldwide, e.g. USA’s BRAIN initiative and Europe’s Human Brain Project, highlight the importance to understand the types of cells in nervous systems. Current reconstruction techniques (both manual and automated) show tremendous variability in the quality and completeness of the resulting morphology. Yet, building a large library of high quality 3D cell morphologies is essential to comprehensively cataloging the types of cells in a nervous system. Furthermore, enabling comparisons of neuron morphologies across species will provide additional sources of insight into neural function.  Automated reconstruction of neuron morphology has been studied by many research groups. Methods including fitting tubes or other geometrical elements, ray casting, spanning tree, shortest paths, deformable curves, pruning, etc., have been proposed. Commercial software packages such as Neurolucida also start to include some of the automated neuron reconstruction methods. The DIADEM Challenge ( <http://diademchallenge.org/> ), a worldwide neuron reconstruction contest, was organized in 2010 by several major institutions as a way to stimulate progress and attract new computational researchers to join the technology development community.  A new effort, called BigNeuron (<http://www.alleninstitute.org/bigneuron>) aims to bring the latest automated neuron morphology reconstruction algorithms to bear on large image stacks from around the world.  The second use story would entail deploying Vaa3D ([www.vaa3d.org](http://www.vaa3d.org)) as an additional service to the active repository described in Use Story I. Vaa3D is open source and provides a plugin architecture into which any type of neuron reconstruction algorithm can be adapted. The second use story would require additional computational resources (and could benefit from multithreaded and parallel compute resources) for the reconstruction process.  In this use story, a neuroscientist user would provide via a web service input parameters to a Vaa3D instance which would trace any recognized neuron structures using a selected algorithm. The output file would be returned via the webservice. | | |
| Expectations *(please describe your expectations for the new technology to be provided by the* ***Open Data Cloud****)* | | |
| Use Story 1 would require:   * A multi-terabyte storage capacity. Each image will typically range from 1-10TB. * A compute node with fast IO bandwidth to storage device (to be specified shortly) * The ability to deploy a Python-based service (BBIC, see appendix I) and supporting libraries (HDF5, etc). * High performance internet connectivity for web service * A standardized authentication/authorization/identity mechanism (first version could provide public access, current version uses HBP AAI). * Web client code for interactively viewing dataset via BBIC service (provided by HBP) * Modern web client (Chrome/Safari/Firefox) for interactive 2D/3D viewing using WebGL and/or OpenLayers. * Sample neuroscience-based volumetric datasets including electron microscopy, light microscopy, two-photon imaging, light sheet microscopy, etc ranging from subcellular to whole brain – provided by HBP, OpenConnectome, Allen Institute and others.   Use Story 2 would require:   * The active repository developed for Use Story I. * The additional deployment of Vaa3D adapted for use with BBIC. A beta version of this is currently available. The REST API may need development. * A multiprocessor compute node with high speed access to the storage device. * Additional datasets including image stacks/volumes of clearly labeled single or multiple neurons – provided by HBP, Allen Institute and others. | | |
| Impacts and Benefits (*Please be specific and use quantified indicators and targets wherever possible*) | | |
| Use Story 1 would enable a neuroscientist user to deploy their data in a specified repository where it would be accessible for web-based viewing and annotation.  Use Story 2 would enable the building a large library of high quality 3D cell morphologies which is essential to comprehensively cataloging the types of cells in a nervous system. Furthermore, enabling comparisons of neuron morphologies across species will provide additional sources of insight into neural function. | | |
| *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:* 5-6 |
| *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:* **equivalent to number of site providing the service** * *Advertisement at events/conferences/workshops:* **4/5 per year** |
| *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:* **10 after the first year of operation** * *Number of conference papers and presentations*: **5 per year** |
| 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)* | | |
| 1. Data repository service for researcher community. 2. Given enough community momentum, as the copyright holder of the software, we could license the image service to imaging equipment companies to allow inter-operability. | | |

**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: Brain Scan Creation**  **Actors**:  **Brain Researcher** working in a group/project performing brain scans. They work for brain research facilities around the world.  **Data Manager** working at Active Repository Center, responsible for maintaining data.  **Action:**  A Scientist is creating a brain scan, which is stored in a form of files. Then the files should be transferred to one of Active Repository Centers and register as well in the central metadata repository.  Some metadata are included in the file but most of them are stored in JSON and XML file. The metadata are important for finding the right scan in the global metadata repository.  Current metadata of scans are: resolution, species, size of the file, number, etc.  Scans are stored in a form of: series of bitmaps,VTK (for 3d rendering),  HDF5, TIFF/JPEG at origin,convert to HDF5 From the data structure point of view a single scan is either file or a directory of files.  For the purpose of processing the scans are converted into HDF5 file and then are transferred to Active Repository Centers and then registered in Metadata Brain Center.  **Current Solution:**  The images are transferred manually to Active Repository Centers where they will be  processed later. In many cases FTP protocol is involved. Brain Researchers upload the scans to ftp server.  In other cases Brain Researchers give access to the data on their own FTP servers to be downloaded by Data Manager, or in some worst case of the largest data sets scenarios Brain Researches send hard drives with data to be uploaded.  **Problems to be solved related to UC1:**   * Data flow from Brain Research Facilities to Active Repository Centers. * Selection of Active Repository Centers to which center the scan should be delivered. There is plan to build multiple processing centers in the world, ideally one per country or more. * Capacity management, how to maintain grants for storage for different groups of scientists. * How to replicate data between Active Repository Centers.--> replication is needed when latency is too high eg. China or in some cases west coast in the US   when replicating data, the location of the replicas will be stored in the metadata.   * Metadata management support * Authentication scheme and authorization scheme for data submitter   **Output:**   * Sets of directories containing brain scans described by some metadata files in JSON or XML formats   **UC2: Remote interactive multiresolution visualization of large volumetric datasets**  **Actors**:  **Brain Researcher** there are brain researchers interested in navigating through existing brain scans.  **Actions:**  Brain Researchers navigate around the brain image using web browser. To make that possible first the Brain Researchers must find a respective scan using central metadata server - this is possible thanks to Metadata Brain Center. Then having direct link to Active Repository Center Brain Researcher is able to dive virtually into the brain thanks to the WebGL standard (→ not hard requirement, graphics acc. not required). But to make that possible Active Repository Center needs to deploy interface which produces WebGL data based on the actual registered image scans. The interface processes data using POSIX api and produces singigicanlty small amout of data in form of HTML+WebGL.  The process of generation WebGL data is IO intensive. The experiments showed that one physical server is able to handle max 10 simultaneous viewers. Expected number of simultaneous users still unknown, want to understand how to scale.  → dynamic load balancing of clients in case of many users accessing the same data set of interest, but this should not be a problem in the short term (2 years)  **Current location of the data:**  CINECA, Juelich, Oslo, SuperComputer Center in Switzerland, UPM Spain  Designated repository per geographical location (the closest).  Current size of data collections is: xxx  **Problems to be solved related to UC2:**   * how to store the data, keep them available through high throughput POSIX interface but still having federated management functionality * access control to data integrated with AAI based on OpenID (→ to start with AAI can be avoided with open data where auth/authz are not necessary) * how to maintain data space (space allocation, cpu utilization etc.) * how to easily deploy Active Repository Centers to make them as many as possible * how to distribute software for active repository centers. In other words releasing new software for image navigation should lead to updating gracefully all the Active Repository Centers (→ there will be many data producers around the world, intention is to provide easy to use means for them to upload the data) * if the cloud will be a solution what would be the cost model. (depend on which cloud and which cloud model. ) * Storage QoS taking into account UC3, which might degrade storage performance for UC2 keeping in mind that UC2 is more interactive and UC3 is more batch processing.   **UC3: Feature extraction and analysis of large volumetric datasets**  **Actors:**  **Neuroscientist** - an actor trying to generate new data based on existing brain scans and register new data into the central metadata server. (Brain researcher is more general term. They have similar access behaviours and access right)  **Actions:**  From the technical point of view in this usecase neuroscientists process data directly from the repositories and generates new data which should be registered in the global metadata server.  **Detailed Description:**  This use case would entail deploying Vaa3D (www.vaa3d.org) as an additional service to the active repository described in Use Case I. Vaa3D is open source and provides a plugin architecture into which any type of neuron reconstruction algorithm can be adapted. The second use case would require additional computational resources (and could benefit from multithreaded and parallel compute resources) for the reconstruction process.  In this use case, a neuroscientist user would provide via a web service input parameters to a Vaa3D instance which would trace any recognized neuron structures using a selected algorithm. The output file would be returned via the webservice.  Output is small and can be transferred on REST interface.  No specific hw needed, 4 cores  **Problems:**   * How to give neuroscientists access to the actual data and provide them in the same time possibility to generate new data. The generated data needs to be stored somewhere keeping in mind limited access and yet registered in the central metadata. * How to efficiently process the data on cloud or grid infrastructure   data is small  **Use case requirements:**   * The active repository developed for UC2. * The additional deployment of Vaa3D ([www.vaa3d.com](http://www.vaa3d.com) ) adapted for use with BBIC. A beta version of this is currently available. The REST API may need development. * A multiprocessor compute node with high speed access to the storage device. * Additional datasets including image stacks/volumes of clearly labeled single or multiple neurons - provided by HBP, Allen Institute and others. (data can be processed at each place, not issue for migration) * Scalable computing environment.   **Output:**  Extracted data objects ….  keeping full control of the access rights to data owners.  **UC4: Publication and citation of data**  **Actors:**  **Data Owner:** neuroscientists having administrative access to the data  **Actions:**  Data Owners should be able to generate persistent citable links to data. It should be work as well with DOI. However by generating citable links data owner will need to take some responsibility to keep that data in the longer perspective. Data owner can define access level to the cited data including anonymous level, by that granting access to anyone having the link. In other cases access control on the cited data should be still possible.  Access to the "cited data" should be monitored to gain ability to generate some data access statistics.  In this use case any object could be cited including: original data sets, subvolumes, extracted object, etc.  A challenging issue in this use case might be ability to cite data being an output of a processing services. For instance an actor wants to cite a part of brain scan which is accessible via image navigation services but not necessarily to copy that data.  **Problems:**   * integration of high throughput data repositories with concepts of persistence and citation of data * limited access to the features of long term data preservations * forking of data between data owners to overtake responsibilities of long data preservation * citing data that have no physical representation but being output of other processing services without storing that. However to achieve long term persistence it might be required to copy the cited data to avoid future problems when the processing service changes and might generate different outputs. * data deletion policy   **UC5: Management of Access Control Rights**  **Actors:**  **Data Owner:** neuroscientists having administrative access to the data  **Actions:**  Human Brain Project maintains its own LDAP repository of users and groups. Based on that there is an OpenID provider which supports identification of the users.  There must be an easy way to maintain ACL rights based on groups membership. Each scan or data object could have an individual ACL rights maintained by the users having respective permissions.  → at the moment it is only accessible by authenticated users, including the user processing off-line  → data owners decide the access controls  **Requirements:**   * System should be compatible with OpenID * ACL |

|  |  |
| --- | --- |
| **Information approved by** | Catherine Zwahlen |

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*) | |
| * Represents in vivo, in vitro and in silico **entities** * Represents **datasets** * Describes properties using **ontologies** * Records **where** an entity or observation is located * Tracks **how** data is produced * Tracks **who** performed experiments/manipulations | |
| **Data size** *(typical size of single file or object)* | Each image will typically range from 1-10TB |
| **Data collection size** *(estimate of total size of data collection in community)* | O(10PB) currently—will grow to O(1000PB) within next 5-10 years |
| **Data format**  *(e.g. XML, CSV)* | Brian scans are stored in a form of: series of bitmaps,VTK (for 3d rendering),  HDF5, TIFF/JPEG at origin, convert to HDF5 From the data structure point of view a single scan is either file or a directory of files. |
| Data Identifiers *(how is the data objects/files identified)* | Current system has index to data and to metadata, searching facilities are provided allowing data discovery. Each dataset is associated with a global unique identifier and there are references (URIs) for multiple representations. For example, there is an entry for a unique dataset and its replica URIs, they link to a common GUID in the metadata system. |
| Standards in use (e.g. FITS, DICOM) | W3C PROV-DM with HBP extensions |
| Data locations (&contacts) | CINECA, Juelich, Oslo, SuperComputer Center in Switzerland, UPM Spain (contacts TBD). |
| Data management plan *(How long should the data be preserved? When can it be made public?)* | * Tier 0 - Unrestricted   + All metadata and/or data freely available (includes contributor, specimen details, methods/ protocols, data type, access URL)   + Reward: Potential citation, collaboration * Tier 1 - Restricted use   + Data available for restricted use, developing analysis algorithms   + Reward: Data citation * Tier 2 - Restricted Use   + Data available for restricted use, nonconflicting research questions   + Reward: Co-authorship * Tier 3 - Restricted use   + Full data available for collaborative investigation, joint research questions   + Reward: Collaboration, co-authorship |
| **Privacy policy** *(Who can access the data?)* | Data use agreement  • Sets conditions for data use  • Don’t abuse privileges (e.g. deidentify human data)  • Don’t redistribute, go to approved repository for registered access (maintain data integrity, tracking accesses)  • Agree to acceptable use policies (e.g. investigate non-embargoed questions only)  • Embargo duration  • Share and share-alike - if data combined with others that should be shared, result should be shared  • Commercial use?  • Stakeholders  • Points to research registry for dataset  • Owners can reserve (embargo) data access for specific research questions for limited time period  • Others may access for non-embargoed use |
| Other aspects | **N/A** |
| *Future Requirements* | |
| **N/A** | |

**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*) | |
| The metadata are important for finding the right scan in the global metadata repository. Current metadata of scans are: resolution, species, size of the file, number, etc. | |
| Metadata Identifiers | **See attached HBP-CORE schema (attached)** |
| Metadata size | < 100K per image stack |
| Metadata format | Some metadata are included in the file but most of them are stored in JSON and XML file. |
| Standards in use | JSON & XML |
| Metadata generation | Metadata are generated when brain scan images are made |
| Metadata locations (&contacts) | Metadata Brain Center |
| Other aspects | **N/A** |
| *Future Requirements* | |
| **N/A** | |

**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.*) |
| * Ingestion:   + Register unique identifiers for each contributor, specimen type, methods/protocols, data types, location, etc   + Mapping metadata for data objects to common HBP data model with provenance info   + Issuing persistent identifiers for data objects in each repository • Data registration REST-API   + Metadata harvesting     - Defining OAI-PMH with common HBP Core data model   + Add entry to KnowledgeGraph – semantic provenance graph * Curation   + Registering spatial data to common spatial coordinates   + Data feature extraction/quality checks   + Update KnowledgeSpace Ontologies   + Augmenting ontologies for metadata (methods/ protocols/specimens, etc)   + Review article defining concepts w/data links * Search   + Indexing to enable discovery of related (integrable) data * Access |
| *Future Requirements* |
| **N/A** |

|  |  |
| --- | --- |
| **Information approved by** | **Catherine Zwahlen** |

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*) | |
| https://docs.google.com/drawings/u/1/d/sMMClcra5qFbD14LbLjYrww/image?w=624&h=688&rev=15&ac=1  **Brain Research Facility**. There are large amounts of image stacks or volumetric data are produced daily at brain research sites around the world. This includes human brain imaging data in clinics, connectome data in research studies, whole brain imaging with light-sheet microscopy and tissue clearing methods or micro-optical sectioning techniques, two-photon imaging,  array tomography, and electron beam microscopy.  **Active Repository Center.** The actual data from Brain Research Facility must be stored in data centers having respective storage solutions plus capacity for data processing.  **Metadata Brain Center.** Maintain the central HBP metadata repository. The repository is a kind of directory for all the available data in Active Repositories and it maintains as well LDAP of HBP members and OpenID interface for authentication. | |
| Data management (Please describe how you access and manage your data sets) | |
| **Community data access protocols** *(e.g. POSIX, GridFTP, WebDAV)* | The images are transferred manually to Active Repository Centers where they will be processed later. In many cases FTP protocol is involved. Brain Researchers upload the scans to ftp server. |
| **Data management technology** *(Please describe what is the data management system in your community, e.g. LFC, iRODS, etc.)* | BBIC (Blue Brain Image Container) (refere to [HBP Volume Imaging Service.docx](https://documents.egi.eu/public/RetrieveFile?docid=2468&version=1&filename=HBP%20Volume%20Imaging%20Service.docx)), which manages 2 data types:   * Stacks of images * 3D volumes |
| **Data access control** *(e.g. POSIX filesystem rights, ACL)* | The interface processes data using POSIX api and produces significantly small amount of data in form of HTML+WebGL. |
| **Public data access protocol** *(How should the data be accessed by public users? e.g. HTTP)* | HTTP queries |
| **Public authentication mechanism** *(e.g. anonymous access, track who downloaded file based on X.509 certs)* | **Data Owner:** neuroscientists having administrative access to the data  **Current status:** Human Brain Project maintains its own LDAP repository of users and groups. Based on that there is an OpenID provider which supports identification of the users.  There must be an easy way to maintain ACL rights based on groups membership. Each scan or data object could have an individual ACL rights maintained by the users having respective permissions.  → at the moment it is only accessible by authenticated users, including the user processing off-line  → data owners decide the access controls  **Requirements:**   * System should be compatible with OpenID * ACL |
| Computing capacities *(Please describe the type and capacities of current physical devices used for your data processing)* | |
| CPU | quad-core minimum |
| GPU | N/A |
| RAM | 8G |
| Storage *e.g., HDD, tapes* | SDD preferred to HDD if necessary. Also may utilize network filesystem storage, but under the network constraints listed below. |
| Network | Fast IO bandwidth, 1GB/s per node |
| e-Infrastructure, *e.g., Clusters, Grid, Cloud, Supercomputing resources* | Cloud, as this is a cloud capability test. |
| Client, *e.g., workstation, desktop, laptop, Mobile device, etc.* | Workstation, laptop, mobile device |
| *Other aspects* | N/A |
| *Future requirements* | |
| **N/A** | |

### 

**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 | process of migration data between active repositories if needed, that might be limited (→ next step requirement) |
| Accessibility |  |  | Y | Simplicity of the access control (UC5) - not that important at the very beginning. |
| Throughput | Y |  |  | See A.3.1 |
| Response time | Y |  |  | See A.3.1 |
| Security |  |  | Y | AAI provided by HBP, data is non-critical |
| Utility |  | Y |  | How difficult distribution/upgrading of the application software onto the environment will be. Docker approach is under HBP investigation now. |
| Reliability |  |  | Y | Support for persistent identification (DOI and persistence life cycle approaches) (UC4) (→ in long-term will be important) |
| Scalability |  | Y |  | Scalability of processing load distribution/brokering, it should be easy (maybe automatic), instantiation of new processing units based on the traffic to active repositories. (UC2) |
| Efficiency | Y |  |  | Simplicity of the process transferring data to the active repository site |
| Disaster recovery |  |  | Y | For the testing phase, the primary location of the data will not be on the EGI infrastructure. During the production phase, data integrity guarantees will need to be defined with the customer community. |
| ***Others performance requirements*** | | | | |
| Effectiveness | Y |  |  | Simplicity of bringing up another active repository |
| Flexibility | Y |  |  | Flexibility of accessing the same data by multiple scientists including intra-groups access |
| Decentralisaion | Y |  |  | Decentralization of resource management. There are many collaborating groups but they remain independent. The system should be flexible to allow independently gain resources by those groups and still provide some integration level. |

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

|  |  |
| --- | --- |
| Software/ applications/services | * *Describe the software/applications/services name, version:*    + BBIC (Blue Brain Image Container, a collection of REST services for providing imaging and meta-information extracted from BBIC files)   + Vaa3D ([www.vaa3d.com](http://www.vaa3d.com)), which provides a plugin architecture into which any type of neuron reconstruction algorithm can be adapted. * *Describe the software licensing:*Open sources * *Describe the configuration:* **will be handled by HBP** * *Describe the dependencies needed to run the application, indicating origin and requirements:* **will be handled by HBP** |
| Operating system | Any modern Linux (RedHat, Ubuntu distribution have been tested) |
| Runtime libraries/APIs *(e.g., Java, C++, Python, etc.)* | * BBIC is implemented in Python use [Tornado](http://www.tornadoweb.org/) web server; an alternative version use Python’s minimalist [SimpleHTTPServer](https://docs.python.org/2/library/simplehttpserver.html) * REST API * HCL (Hyperdimensional Compression Library) |
| Typical processing time | 1 day for 4TB image stack, extrapolated from current small scale tests |

**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.).* |
| Requirements for Infrastructure Providers:  * **PERFORMANCE**. Being able to work at throughput 1GB/s per node to support 10 users (for image service nodes, throughput between the data server and the storage) * **STORAGE CAPABILITIES**   + - HDF5 files at size up to 10TB     - Posix access to data * **CO-LOCATION**. cpu close to data |

**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* |
| * Docker approach is currently investigated in HBP * iRODS is not preferable. |
| *Does the user have preferences on specific resource providers? (e.g. in certain countries, regions or sites)* |
| Centres with access to PRACE backbones would be good for testing. |
| *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)* |
| irrelevant |
| *Does the user (or those he/she represents) have access to a Certification Authority? (to obtain an EGI certificate)* |
| Yes |
| *Does the user need access to an existing allocation (🡪 join existing VO), or does he/she needs a new allocation? (🡪 create a new VO)* |
| TBD. |
| *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)* |
| Not currently evaluated |

|  |  |
| --- | --- |
| **Information approved by** | Jeff Muller |

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