**Document Title:** HBP Use Cases

**About:** Requirements extraction from the EGI-HBP meeting, 21 April

**Version**: Draft

**Date:** 6 May 2015

**Status:** To be approved by HBP

**Use cases**

**1)** HBP is developing an image service (named?) which manages all issues that providing multi-rang solutions of the use onto the data, searching of sub-volume of data, also be able to do arbitrary facing angles. This is a basic service for core processing like to register data and align data to standard template brain spaces. The service is developed using Python, HDF5. Data container is currently plug into several web clients using desktop analysis software.

**2)** HBP plans to deploy an analytical prices of software that runs either on a single thread or multiple threads to analyse, for example, to extract neuron morphology, which needs a tracing algorithm for large collaborative data neutrons which is recently launched a world-wide efforts of collection of large neuron image sets. The aim is to use automatic reconstruct algorithm to trace these neutrons and extract the object that from the image sets which them be analysed by cluster algorithm. This may require to run ideal in parallel reconstruction algorithms and be able to compare the results.

HBP is looking for different type of repositories that not just create a collection of large images that have to be transferred. The intention is to active data services for accessing just sub-portion or resolution of that data wanted and ultimate be able to do analysis where that data sets are. The purpose is to leave the data in place.

**Computation Facilities:**

Imagine processing facilities include: imagine service that running operational in super computer centres; web clients operational for 2D/3D; desktop workstation platform with analysis program and interface to the service operations. Request for Core 15G RAM, high I/O storage as minimal configurations. They will collect information about standard use load.

**Image data size**: current image data sets are around 1~10 TB. With the quick growth of the imaging technology, the data collection can be grown quickly. E.g. in China it makes commercial for scanning. However, HBP does not expect their data grow will exceeding the nowadays computing capabilities.

**I/O (file store to service)**:1GB/s per node under low testing condition

**Replications**:

Currently data are not replicated. However for optimising the performance, they will consider to replicate some key datasets to sites where lower latency is desirable. For example, there is a use case, where they are looking for to replicate large datasets from China to Europe.

**Data Access**:

Typically users will access data through the browsers. But they also have desktop clients with multi-resolution volume neurone analysis functions. There is a need to provide tool/service for data producers to upload data to the repository, high performance protocol for data transferring will be preferable to get those data to the repository effectively. There is a need to support all different types of data producers, e.g. to provide one default efficient transferring services, and a set of recommended tools/services. Currently, large data producers are those associated with super computer centres; others may have lab service or university server but don’t necessary have e-infrastructures to sustain the in a long-term, there are also use cases e.g., a group in China would like to make their data more widely available.

* Data are open. But some data will need for AAI, e.g. some data producers want to limited their data access for collaborations. In general there is no identifiable individual human data.
* Data access pattern: Depending of different use cases, arbitrary facing angle will need random data access, but many computation, e.g., for quick computed structure and quick computed projection are serial/sequential data access.

**Architecture**:

Centralised data archive within one data centre. The intention is to provide a long-term/permanent data access point.

**Dataflow**:

Data producers transfer data to the central data archive, data would be accessible without having to replicate the whole dataset but having the level of the details what they would be interested or the sub-portion that necessary for the analysis. Local data analysis workflow will need data injection at different stages. Raw data can be precompute and will be directly stored in the HDF5 files, but in general, many of them will be generate on the fly and not stored. e.g., users just rapidly navigating through the volume and search for a sub-volume of data and project them. However, in the second use case, there will be a need to capture immediate processing objects, since end user’s activities will be very exploratory, they may need to reuse the same results for later uses. Such processing results are not significant in volume, which is in MB level.

Data will be reduce during processing pipeline. There is few cases that processed data will be bigger than the raw datasets.

**Imagine processing**:

Currently, some image processing happen in the image service on the server sites, and the rest happen with web client in the browsers.

**Metadata**:

Some metadata are in HDF5 files referring to the ontology that operating on the ontology services. Other metadata can be searched by the knowledge graph.

**Data Identification**:

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.

One details is the requirement for I/O performance which mentioned in previous conversations: which requests **1GB/s per node.**

**Requirements for Cloud platform**:

There are interesting/plan for CDMI development.

**Questions to HBP**

1. We want to discuss in a bit more details some basic use cases, staring from the origin of the data ending at the processing it. We need to be able to divide it somehow to better understand.
2. We would like to understand what HBP expects from EGI testbed. We would need to know what are the factors of success we want to measure by testing our testbed. What are the most important issues HBP is looking to solve?