CSC – IT CENTER FOR SCIENCE

EUDAT – EGI/ENVRI

EISCAT-3D data pilot

Ari Lukkarinen

10/30/2014

This document describes observations and results of a pilot project aiming to enhance co-operation between EGI and EUDAT in EISCAT project.

1. Introduction

In this joint pilot project between EUDAT (http://www.eudat.eu/) and EGI/ENVRI (http://www.egi.eu/, http://envri.eu/), the purpose was to test EUDAT services with EISCAT (https://www.eiscat.se/) data and EGI services.

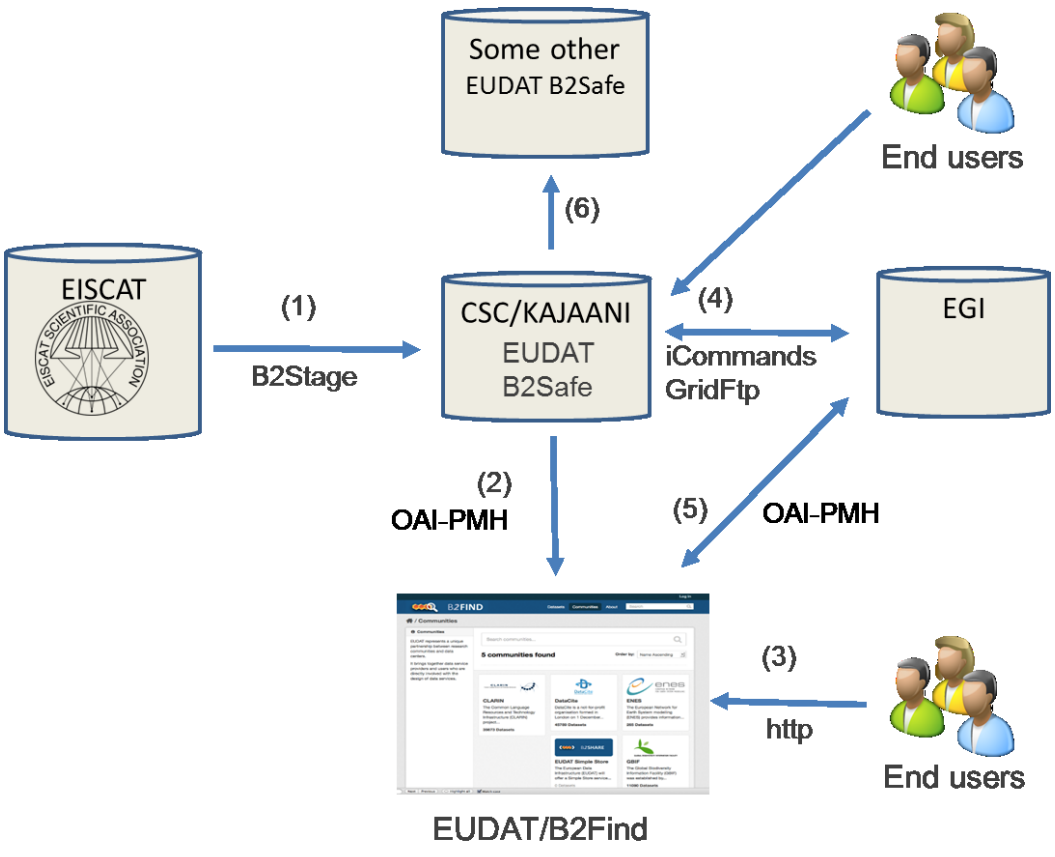
The pilot started at the beginning of 2014. Last data transfer tests were done during August 2014. The size of the dataset was about 1 TB and contained roughly 400 000 files.

This pilot project was done by following persons:

- Ari Lukkarinen (CSC) ari.lukkarinen@csc.fi
- Jani Heikkinen (CSC) jani.heikkinen@csc.fi
- Ville Savolainen (CSC) ville.savolainen@csc.fi
- Pekka Järveläinen (CSC) pekka.jarvelainen@csc.fi
- Ingemar Häggström (EISCAT) ingemar.haggstrom@eiscat.se
- Salvatore Pinto (EGI) salvatore.pinto@egi.eu
- Małgorzata Krakowian (EGI) malgorzata.krakowian@egi.eu

2. Environment

The environment used in the pilot was as follows:



EUDAT B2Safe and B2Find services were used in the pilot. B2Safe service provides reliable storage. Service is based on iRods which is an open-source data management software. It functions independently of storage resources and abstracts data control away from storage devices and device location. More information about these tools can be found from EUDAT B2Safe[1] and iRods[2,] web pages.

EUDAT B2Find is a user-friendly metadata catalogue of research data collections stored in EUDAT data centers and other repositories. Service allows users to find collections of scientific data quickly and easily, irrespective of their origin, discipline or community. B2Find service is based on CKAN. More information can be found from EUDAT B2find and CKAN web pages[3].

EISCAT-EUDAT-EGI environment functions as follows (numbers below refer to numbers in figure above):

(1) As a first step EISCAT data was copied form EISCAT archive to EUDAT B2Safe service through the use of EUDAT B2Stage which was implemented in this case with GridFTP. At CSC, metadata was automatically extracted from HDF5 format data files and stored into iRods service database (iCat catalogue). The database maintains information about data and metadata stored in the iRods service.

(2) iRods service was extended to provide an OAI-PMH interface[4] that is commonly used to harvest metadata from data repositories. This interface was tuned to make the mapping between iCat catalogue and OAI-PMH interface. The metadata was then harvested to B2Find service.

(3) End users were able to use the B2Find service to search information about the data. In the original data there were of the order of 70 different parameters. A subset of these was chosen to be stored to B2Find service as individual attributes. Other parameters were stored as additional metadata information.

(4) Data that was made discoverable through B2Find service was accessible by EGI from B2Safe using either iRods iCommands or GridFTP. A remote command line access from researcher's own computer was possible.

[1] http://www.eudat.eu/b2safe
[2] http://www.irods.org
[3] http://www.eudat.eu/b2find and http://ckan.org/
[4] http://www.openarchives.org/pmh/

3. Results and observations

Metadata extraction and management required some amount of work, but after initial modifications and tunings metadata was successfully extracted to B2Find.

GridFTP based data transfer requiring personal certificates caused some issues, but, after tool, directory and file visibility, and certificate based issues were solved, the service worked without problems.

As a whole, the target – to test EUDAT storage services with EISCAT data – was successfully achieved.

4. Next steps

In case co-operation between EGI, EISCAT, and EUDAT continues, following issues should be considered:

Metadata synchronization ((5) in figure above) between EGI and EUDAT could be enabled in case 1) EGI portal would provide an OAI-PMH interface to metadata and in case EUDAT B2find service would allow metadata transfer from the service. At the moment B2Find service only harvest metadata from external sources.

Data replication between B2Safe services ((6) in figure above)) and PID generation were not implemented. However, both of these tasks are routinely done so that implementing these tasks should be a simple task. PID generation was not implemented, because we were only testing the usability of the service.

There were also discussions about OpenSearch[5] service that B2Find service could offer. However, on EUDAT point of view it seems that although the idea of OpenSearch is solid, it seems that current browsers do not support OpenSearch technology that well. Nevertheless, discussions about the role and function of OpenSearch could be beneficial for EUDAT project. For example, on end user point of view, could a properly designed REST API be used instead of OpenSearch?

[5] http://www.opensearch.org/Home