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.

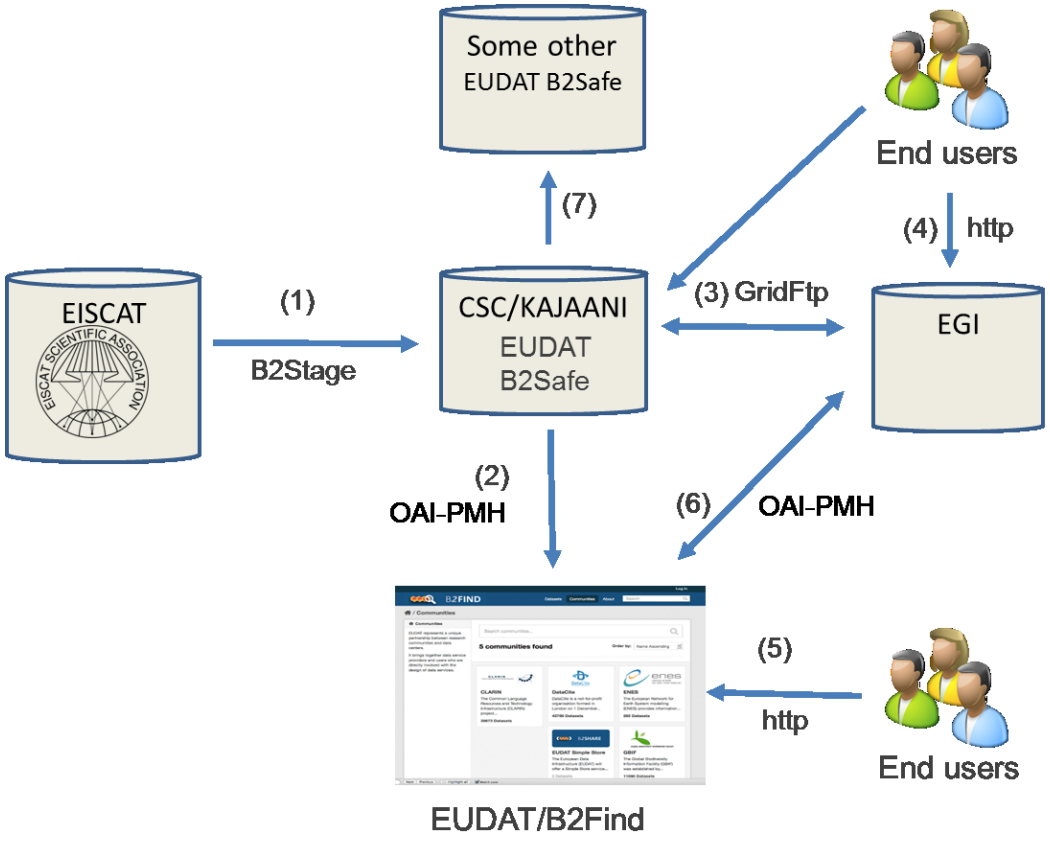
The pilot started at the beginning of 2014. Last data transfer tests were done during August 2014. The size of the dataset was about 1 TB and contained roughly 400 000 files.

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:



From the EUDAT service portfolio B2Safe and B2Find services were used in the pilot. B2Safe provides reliable storage service. 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].

On EGI side ENVRI OpenSource Geospatial Catalogue (OSGC) was used to store the data and the metadata. The data in OSGC service was available via http.

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 iCAT database maintains information about data and metadata stored in the 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.

(3) Data that was stored in B2Safe service was accessible by end via command line access using iCommands or GridFTP.

(4) Data was also available via EGI (web interface via Data Access and Dissemination PaaS based on ENVRI OSGC[5]). Connection between B2Safe and EGI was done via GridFTP.

[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/
[5] For more information about OSGC, see http://sourceforge.net/projects/osgcat/

3. Results and observations

According to initial plan, information about data was planned to be stored to B2find service. End users were planned to use some predefined criteria to search the data from B2Find (5). Metadata extraction from the original data and management required some amount of work, but after initial modifications OAI-PMH interface of the B2Safe service was ready for metadata synchronization. Unfortunately, final transfer to B2Find was not done due to technical difficulties. These problems were most likely due to a mismatch in OAI-PMH attributes between B2Safe and B2Find services or there was some other technical issue that prevented the metadata synchronization.

Another problem that caused issues with metadata, was lack of it. B2Find service requires some predefined metadata elements to be present in the data. Old experimental data, like the old EISCAT data, does not necessarily contain such data. This issue should have been fixed *e.g.* by manually adding information directly to B2Safe metadata catalog.

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 rather well achieved. We were able to move and store the data as planned, but there were some difficulties with metadata.

4. Next steps

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

Metadata synchronization between B2Safe and B2Find (5) should be solved. In addition, metadata transfer between EGI and EUDAT ((6) in figure above) 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. B2Find service description should clearly define what services it provides and what features are not supported.

Data replication between B2Safe services ((7) 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.

5. EUDAT 2 DISCUSSION

On EUDAT point of view, managing repositories with insufficient metadata elements is a generic problem, that should have and generic solution. EUDAT community should define a best practice to solve the issue.

Another generic issue is the number of parameters in the original EISCAT data. The number of parameters was of the order of 70. In this case a subset of these was chosen to be stored to B2Find service as individual attributes. Other parameters were planned to be stored as additional metadata information. Supposing that B2Find will later contain information about data managed by EUDAT and metadata large of number of scientific communities, what metadata should be stored in B2Find/CKAN database as individual attributes and what metadata should be managed by some other means? Number of experiments and, therefore, number of metadata attributes is so large that everything cannot be stored in a simple relational database.

There were also discussions about OpenSearch[6] service that B2Find service could offer. Discussions about the role and function of OpenSearch could be beneficial for EUDAT project. For example, on end user point of view, does OpenSearch provide some benefits for end users that could and should be provided by a B2Find REST API that has been under consideration?

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