EOSC-hub

Task 6.4: Data and metadata management

Paolo D'Onorio De Meo (CINECA)

09 January 2018

Outline

- Task 6.4 covers the area of **data management**
  - data entities, with provenance and associated meta data are registered as data objects, often bundled as data collections, with provenance information and annotations (metadata)
- It ensures maintenance and integration of data management services

Objectives

- **Integration** and **maintenance**
- Total effort of 104 PM in 36 months

Services

- EUDAT services
  - B2SAFE
  - B2SHARE
  - B2HANDLE
  - B2NOTE
- EGI services
  - *Indigo DataCloud* - OneData
  - *OpenAIRE* - Zenodo
  - *DataCite* services and DOIs

EUDAT Services

B2SAFE

B2SAFE provides an abstraction layer of large scale, heterogeneous **data storages** and guards against data loss in long-term archiving.

It implements **data management policies** (DMP) on their research data across multiple administrative domains in a trustworthy manner.

B2SAFE is the most mature service in the EUDAT catalogue. Its status is production and it reached version 4.0.1.

(Claudio Cacciari - CINECA)

EUDAT Services

B2SHARE

B2SHARE is service towards researchers, scientific communities and citizen scientists to store and **publish** small-scale research data from diverse contexts. Since this task is all about the metadata this is service is fundamental to our activity.

Its status is production and it reached version 2.0.1.

(Mark van de Sanden - SurfSara)

EUDAT Services

B2HANDLE

B2HANDLE is the distributed service for storing, managing and accessing **persistent identifiers** (PIDs) and essential metadata (PID records).

Its status is production and it reached version 1.1.

(Tobias Weigel - DKRZ)

EUDAT Services

B2NOTE

B2NOTE is a service that allows you to easily and intuitively create **annotations** on research data hosted in the EUDAT Collaborative Data Infrastructure.

Its status is "pilot" (not yet reach production level).

Its latest release was 0.5.

(Yann Le Franc - eSDF)

Onedata

The Indigo DataCloud project has this **global data management system** providing easy access to distributed storage resources for data management and data-intensive scientific computations.

Its status is production and it's currently working on releases candidates of the version 3.0.

Zenodo

In the context of the *Openaire* project Zenodo is the repository implementation to share metadata in Open Science.

It is based on Invenio, status is production.

There is no release using Semantic Versioning on its software code repository.

Datacite

DataCite is the organisation that provides persistent identifiers (DOIs) for research data.

Goals: Maintenance

- **Software updates**

  - to the latest releases/patch to avoid security vulnerability

  - and to ensure the latest features available by the underlying technologies

- **Scalability**

  - Hardware and/or software operations to ensure **performance** upgrades during the course of the project

for each service operational within the task 6.4

Goals: Maintenance

- The work completed for any maintenance among the first year will be reported in the Deliverable 6.2 at Month 14.

Goals: Integration

- Integration is clearly the main objective
- especially between services that have been created to operate under the two flags of EUDAT and EGI

Goals: Integration 1

- **Authentication** has the key role to start any integration

- A continuous communication must be opened with the EOSC-hub 5.1 task, on how CheckIn (EGI / Indigo IAM) and B2ACCESS (EUDAT) would be reaching *uniformity*
  - What about OpenAIRE?

- Our task members have to be aware of how the uniformity will be programmed to direct the maintenance of the services towards the same path

Goals: Integration 2

- **Harmonisation** of the **identifiers handlers**
    - EUDAT currently relies its data registration to the Persistent IDentifier (PID) Handle (operated by SurfSara)
    - while EGI bases his data persistence on the Digital Object Identifier (DOI) system

Goals: Integration 3

- Crucial aspect in fusing EGI and EUDAT data into the same hub is the shape of the metadata records

- The **metadata models and schema** of the current datasets have to be adjusted or expanded to be compatible with each other

Goals: Integration 4

- **Metadata exposition**

  - from EUDAT to EGI and vice versa

- B2SHARE and ZENODO already provides that functionality through their common software base *Invenio*, via **OAI-PMH**

- Find out how this will operate against the other services of the data catalogue workflow

Goals: summary for the first year

1. Setup communications
2. Find out if more requirements
3. Maintenance
4. Authentication
5. Harmonisation of Identifiers
6. Metadata schemas
7. Metadata exposure

Cross-WP activity

- In the course of the first year WP 4, 5 and 10 will have some aspects in common

- Documents that have to be taken into account soon:

  - D5.1 Initial maintenance and integration plan for federation and collaboration services (month 4)

  - D4.1 requirements for the services in the catalogue (month 6)

  - D10.1 EOSC Hub Technical Roadmap v1 (month 12)

  - D6.2 First report on WP6 the maintenance and integration of common services (month 14)

Any Questions?