www.egi.eu
@EGI_eInfra

Data Management in EGI

Status of EGI Services for Data Management

Baptiste Grenier
EGI Foundation

The work of the EGI Foundation
is partly funded by the European Commission under H2020 Framework Programme

Outline

- A quick overview of all EGI Data-related services
  - Archive Storage
  - Online Storage
  - Data Transfer
  - DataHub
- Looking forward
  - Offering services for the whole data lifecycle
  - Interesting complementary solutions

EGI Archive Storage

Back-up data for the long term

- **Legacy** service for **long term archival of data**
  - Initially based on Tape
    - Multiple EGI participants using tape for their internal needs
  - Other technologies could be used

- Currently discussing access models with service providers
  - Any interested service provider is welcome
  - May be retired from EGI Service Catalogue if no way to offer it is found

Online storage

Store, share and access your files and metadata

- A **production** service covering multiple **data storage and access** technologies
  - "Grid" Storage Elements
    - DPM, dCache, StoRM…
  - "Cloud" object-storage
    - OpenStack Swift

SWIFT
an OpenStack Community Project

dCache

DPM
Disk Pool Manager

StoRM

- **Providing data access** to other services
  - Cloud Compute, Cloud Container Compute
  - High Throughput Compute
  - Workload Manger

Data Transfer

Transfer large sets of data from one place to another

- A **production** service for **managing data transfers** reliably and efficiently
  - FTS- and WebFTS-based

- Two providers have **recently agreed an OLA** with EGI
  - CERN – FTS and WebFTS
  - UKRI/RAL - FTS

- Next steps
  - **Integration** of a **with EGI Check-in and RCauth**
    - OIDC
    - Online CA

DataHub

Data as a Service, distributed data management

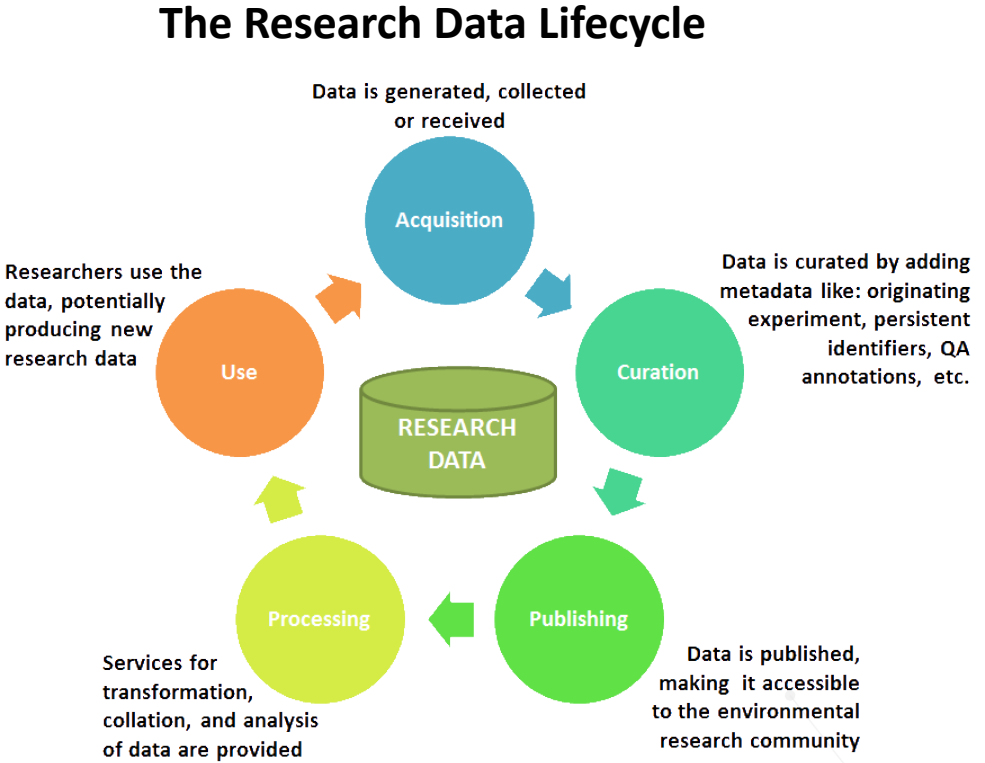
- A service for **federated access, sharing, publishing, discovery of data**
  - based on Onedata, operated by CYFRONET
  - EGI's central Onezone: https://datahub.egi.eu
    - A reference/central Oneprovider connected to the DataHub
    - Providers at various sites can be federated
  - Integration with various other services
    - EGI: Check-in AAI, Cloud Compute, Cloud Container Compute, Online Storage, Notebooks,…
    - EOSC-hub: B2HANDLE, B2FIND,…

- A long exploratory phase with various piloting activities was conducted
  - Lots of improvements on stability, performance
  - Lots of new features

Looking forward

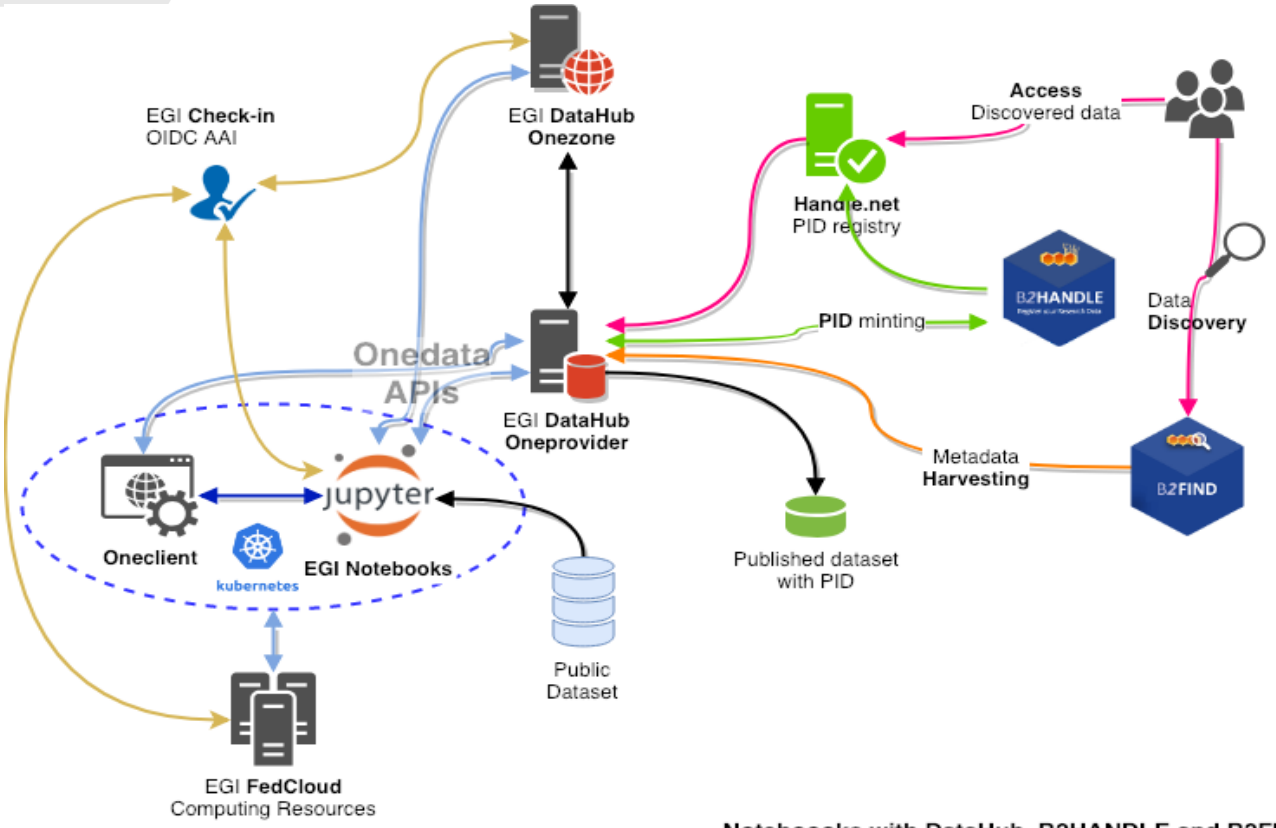
- Ambition: Providing **an interoperable** and **composable service offer** supporting the **research data lifecycle**

- Supporting FAIR practices
- Persistent Identifiers
- Cataloguing, Discovery
- Interoperability, Integration
- Data orchestration



Source: ENVRI Reference Model

Integration example



Notebooks with DataHub, B2HANDLE and B2FIND

Technology scouting

Rucio: distributed data management

- **Rucio**
  - A distributed data management solution
  - A rule-based engine for automated data management
  - FTS for Data Transfer scheduling and management
  - A central service able to connect to existing storage endpoint
    - No new component to be deployed at a site level

Technology scouting

Invenio: building data repositories

- Invenio
  - A **framework** for building data repositories
    - DOI/PID, discoverability, citation
    - Multiple “implementations”: Zenodo, CERN Videos, CERN Open Data, Reana, B2SHARE,…
  - Multiple options
    - Using/adapting/investing effort in an existing implementation
    - Creating a new one targeting EGI needs

EGI: Advanced Computing for Research

www.egi.eu
@EGI_eInfra

Thank you
for your attention.

Questions?

This work by the EGI Foundation
is licensed under a Creative Commons Attribution 4.0 International License.

The work of the EGI Foundation
is partly funded by the European Commission under H2020 Framework Programme