www.egi.eu
@EGI_eInfra

# Federated Data Management

## *Requirements and Roadmap*

**Baptiste Grenier**
*EGI Operations*

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

# About users' requirements

*Looking at a few user communities' needs*

# Data Management Requirements

*From July's Workshop on Data Management*

- Three entities **representing** a broad range of **solutions** and **services providers**




- Multiple **user communities**
  - PaNOSC
  - ASTRON
  - Earth Observation JRC




- Goal: *proof-of-concept and pilot use cases that will be implemented by XDC, ESCAPE and EGI in collaboration with the invited scientific communities*

*https://indico.egi.eu/indico/event/4698/*

# Earth Observation JRC

- The designed e-Infrastructure should
  - provide optimal data **indexing**,
  - **transparent** resource **marshalling**,
  - **smart caching**,
  - massive **parallel** (micro-) **tasking**,
  - vector/raster **integration**,
  - advanced **visualization** support,
  - **open** standards based,
  - **open** source (both server and client side)

# ASTRON Science Data Center

*Use case about LOFAR data*

- The designed e-Infrastructure should facilitate
  - **Finding** data
  - **Staging** data
  - **Processing** data
  - **Analysing** results
  - **Publishing**/sharing results
  - Authenticating using a **single sign-on** mechanism

# PaNOSC use cases

*Photon and Neutron Cluster*

- **Remote archive** of a facility **experimental data**
  - perform analysis on data sets obtained during an experiment
  - **initial** transfer around **600TB**, **annual** volume from **300TB up to 10PB**
  - possibility to **retrieve** data **on-demand**
- **Remote Data analyses** using **Jupyter** notebook services of another provider
  - **compute** and **data** infrastructure are **distant**
  - a service to **host the results**
- User's data transfer to its home organization/home computer
  - The **volume** of data (up to 100TB) is **important**

# Pilots based on multiple solutions

*An ongoing activity*

- Composing RI-specific custom e-infrastructures leveraging different solutions
  - Check-in, dCache, Onedata, Rucio, FTS, INDIGO PaaS Orchestrator,…

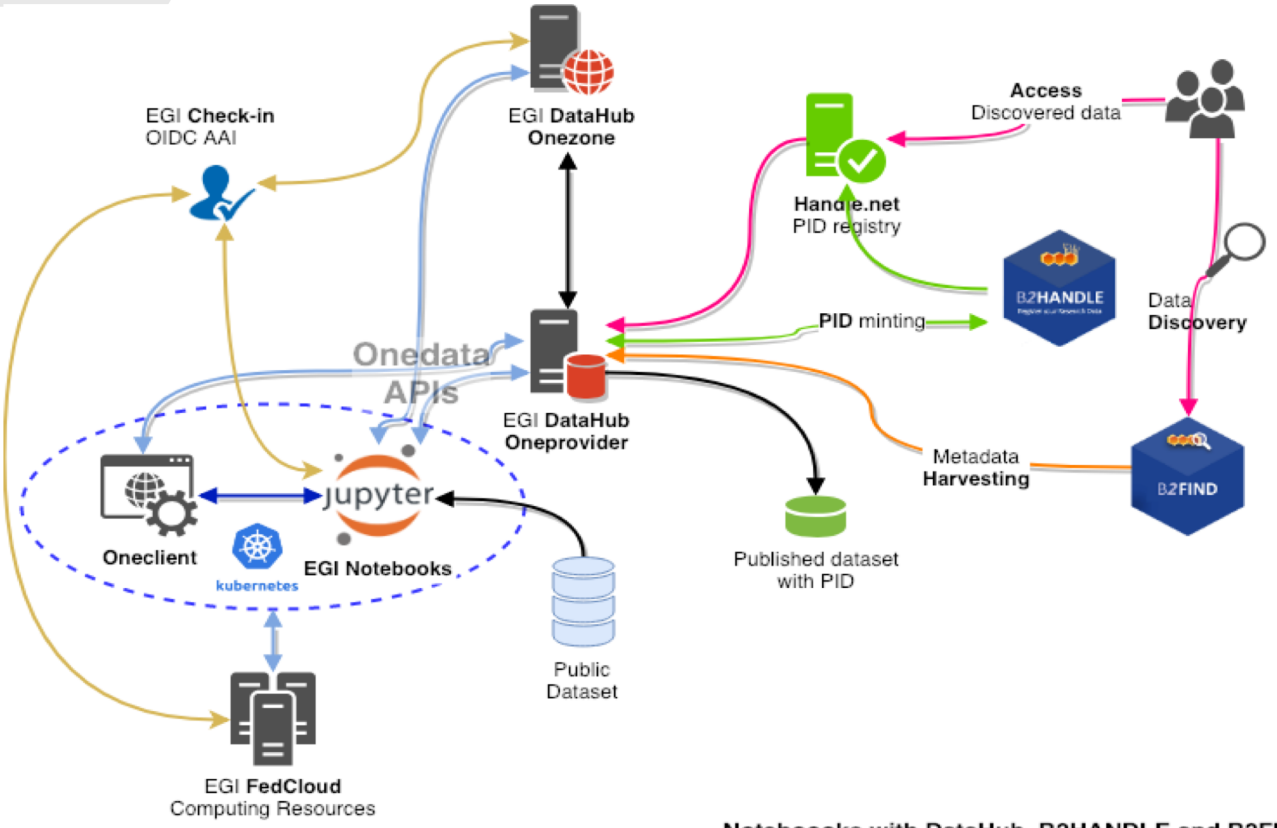
ONEDATA

EGI Check-in service
Secure and user-friendly federated authentication and authorisation

RUCIO

dCache

# Integration activity in the EOSC context




Notebooks with DataHub, B2HANDLE and B2FIND

# EGI Services for Data Management

*Status and next steps*

# EGI Archive Storage

*Archive data for the long term*

- A service for **long term archival of data**
  - Initially based on Tape
    - Multiple EGI participants using tape for their internal needs
  - Open to other technologies and higher-level solutions


- Currently discussing access models with service providers
  - A partnership with CINES is being formalised
  - Additional interested service providers are welcome

# 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

- A stable solution **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 **agreed an OLA** with EGI
  - CERN – FTS and WebFTS
  - UKRI/RAL - FTS

- Next target
  - **Integration with EGI Check-in**
    - OIDC and token translation

# 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

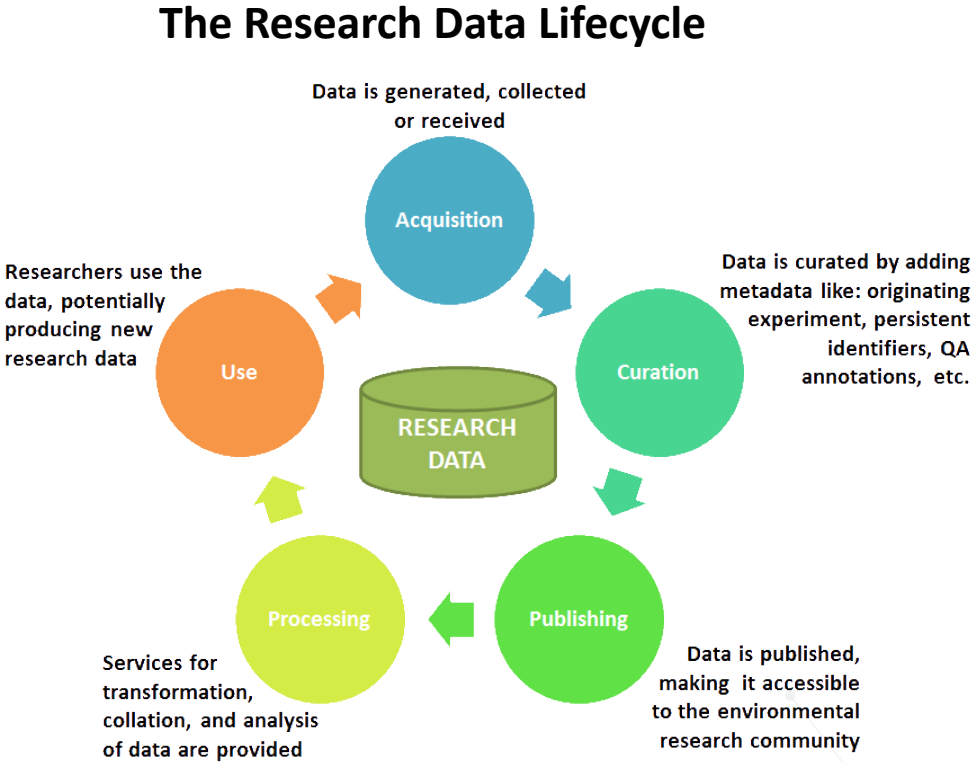
# Supporting additional needs

*A glimpse of the future*

# 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*

# 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,…
  - Looking forward to the outcome of the InvenioRDM project
    - « a turn-key open source research data management platform »

*https://inveniosoftware.org/blog/2019-04-29-rdm/*

# EGI: Advanced Computing for Research

www.egi.eu
@EGI_eInfra

Thank you
for your attention.

*Questions?*



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