# EUDAT

## How manage Data into the Collaborative Data Infrastructure: a general overview of EUDAT services

Giovanni Morelli

www.eudat.eu

EUDAT receives funding from the European Union's Horizon 2020 programme - DG CONNECT e-Infrastructures. Contract No. 654065

# Outline

## What kind of problems we want(try) to solve

- Different management system for different communities
- Quality of data sets
- Class of users

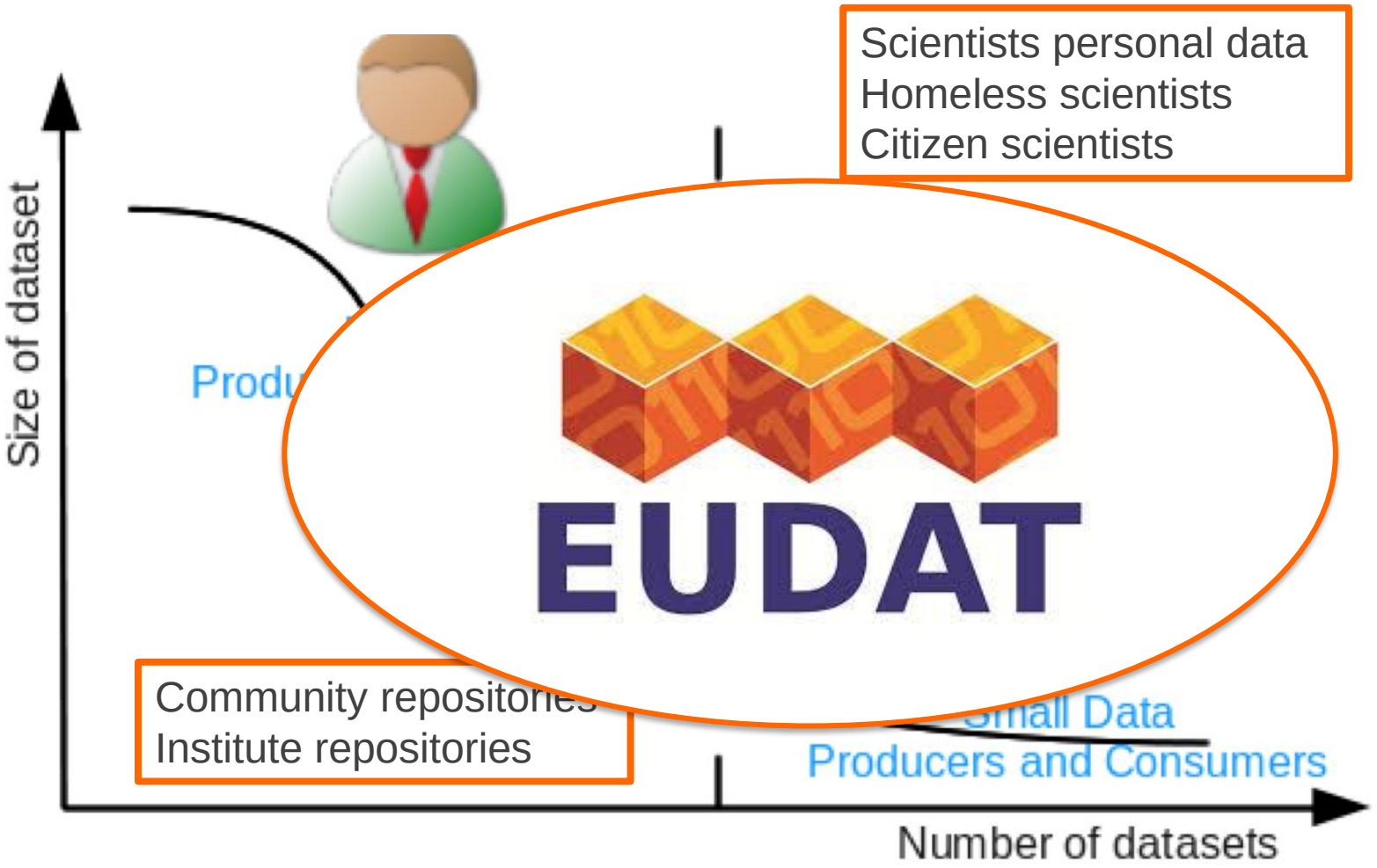
## What about our solutions (B2<services>)

- B2DROP, B2SHARE,B2SAFE,B2STAGE,B2HANDLE,B2ACCESS,…
- B2<service> integration

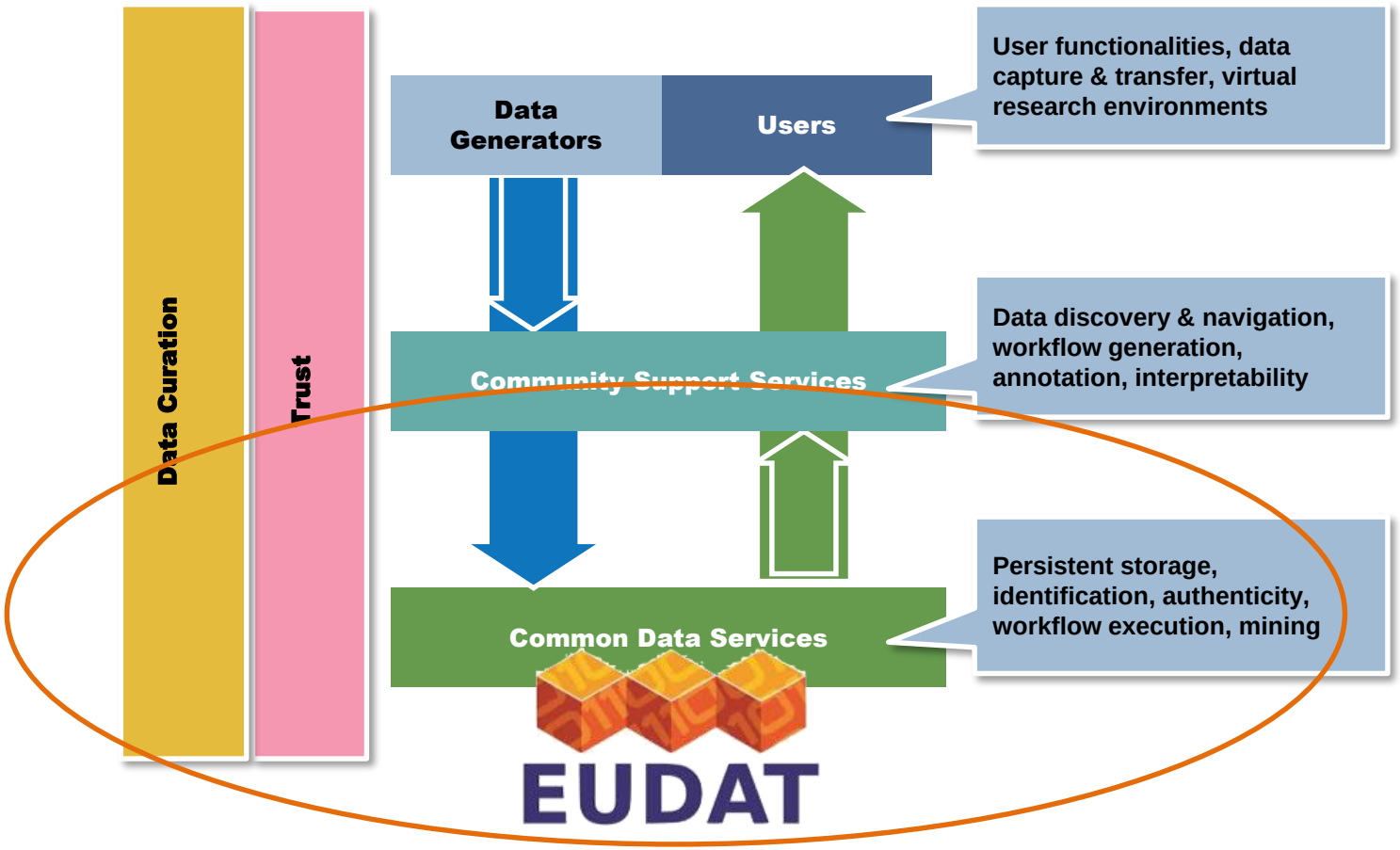
## Project and Service Enabling

- Community / EUDAT interaction
- Practical use cases

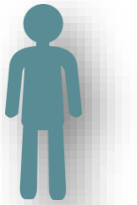
# Where Does EUDAT Fit In?
## (in a Data quality view)



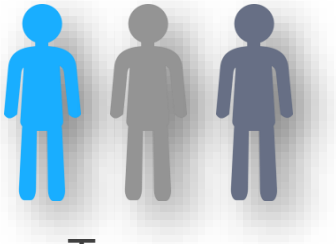
# Where Does EUDAT Fit In?
## (in a multilayer view of Data Management)



# Who can use EUDAT service



Single researcher



Team

Community

Upload and download

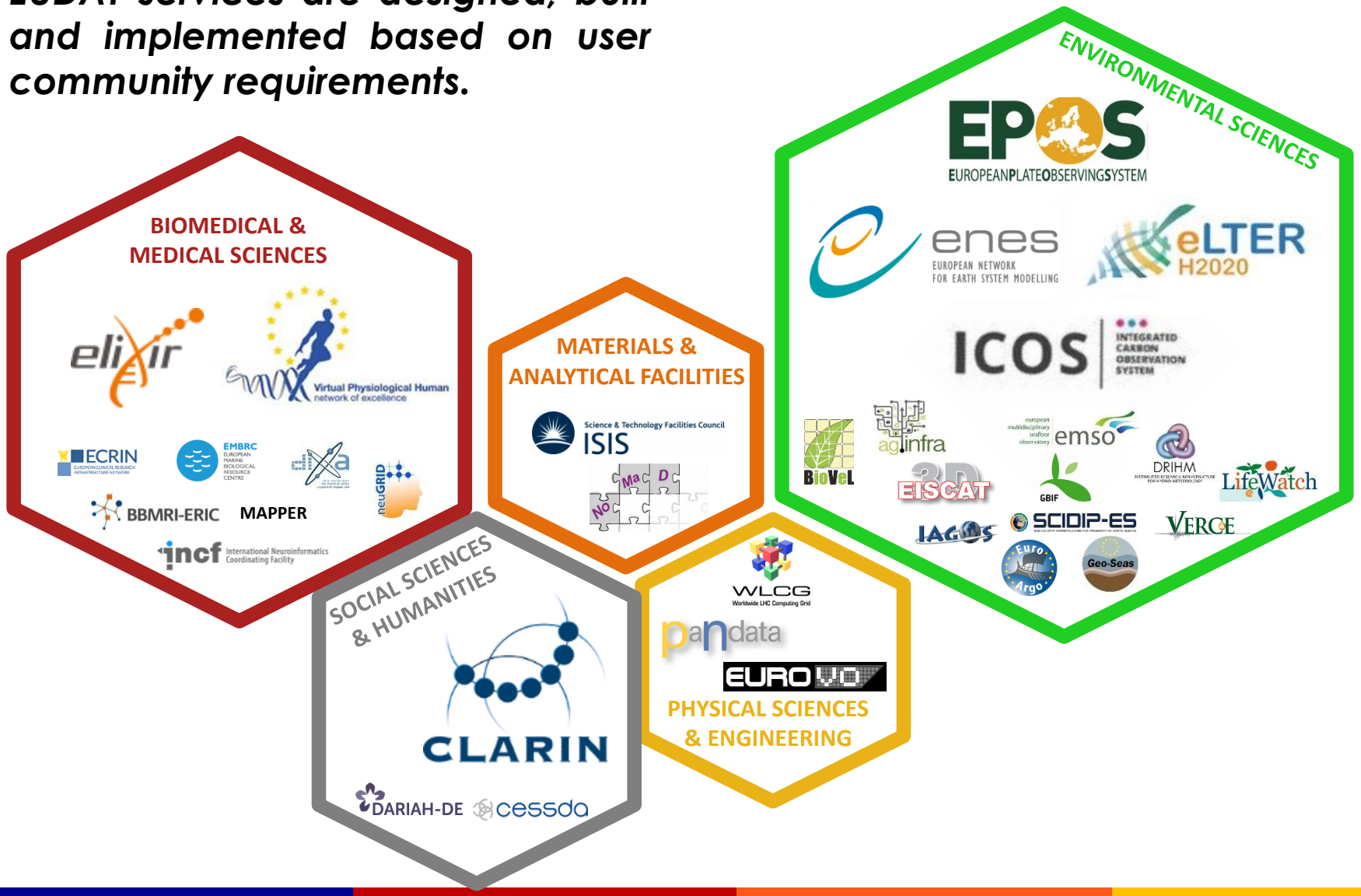
Upload, add metadata, share

Periodic transfers, quality checks …

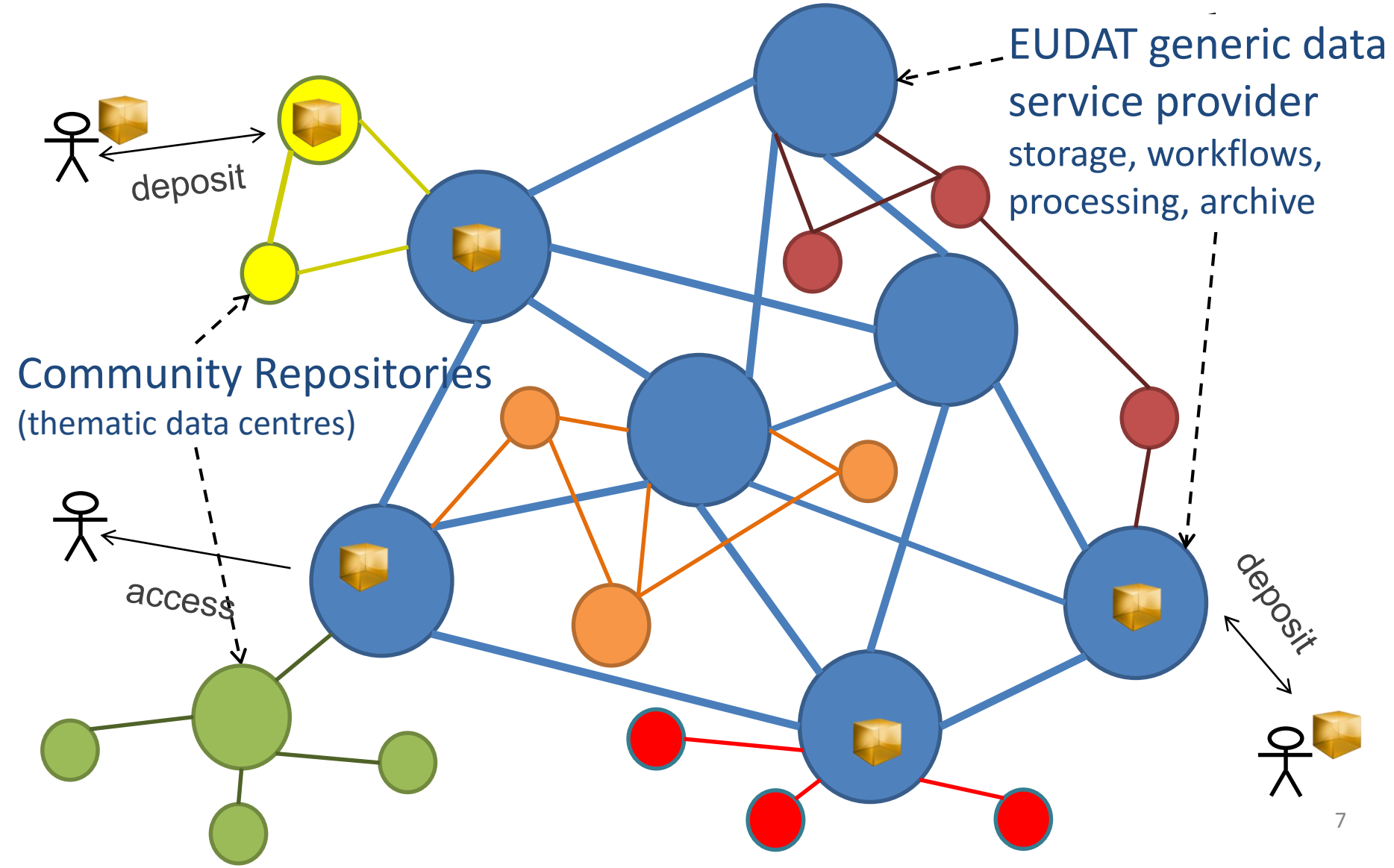
Different strategies for different usage scenarios

# Community-Driven Solutions

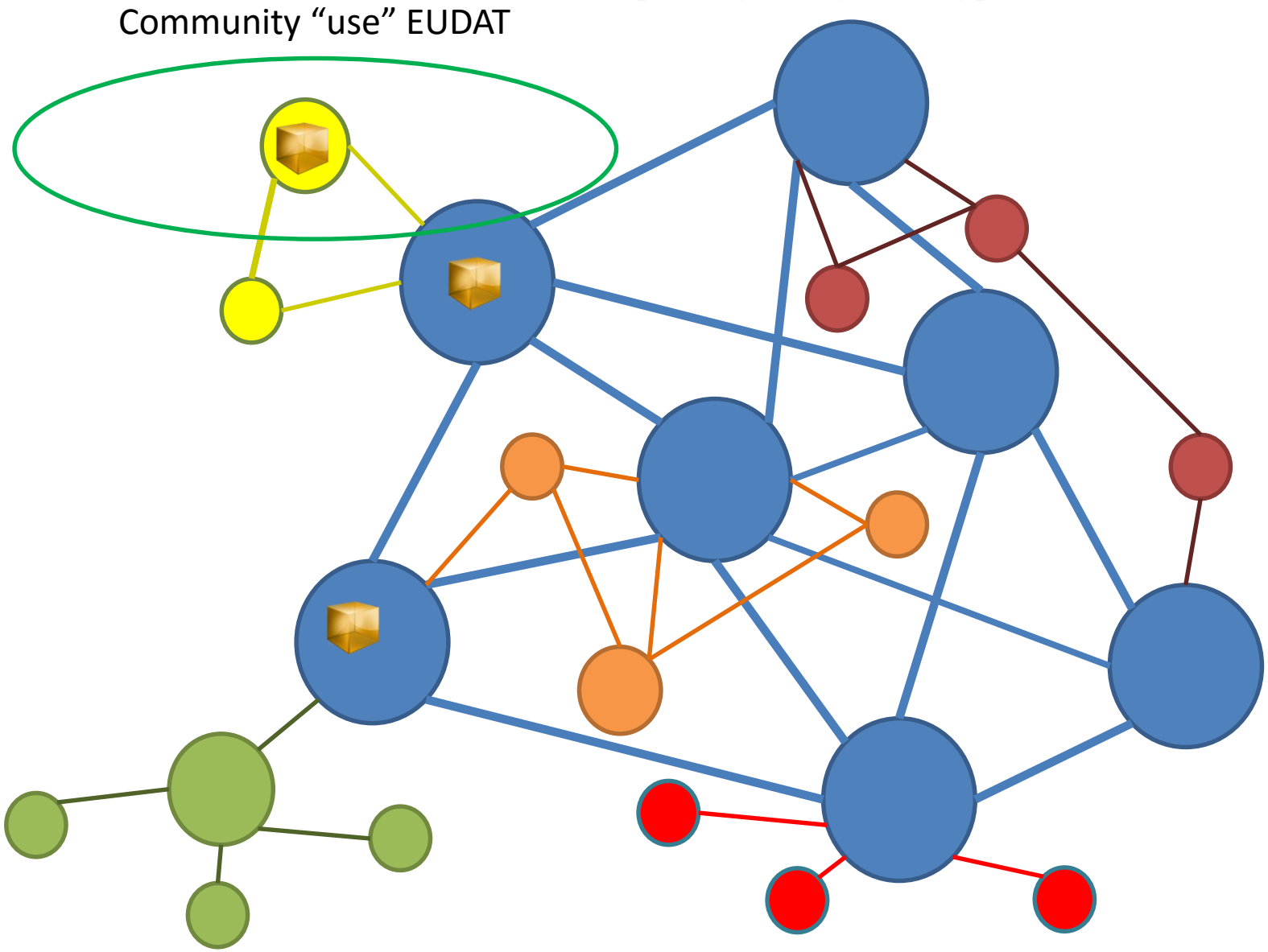
***EUDAT services are designed, built and implemented based on user community requirements.***



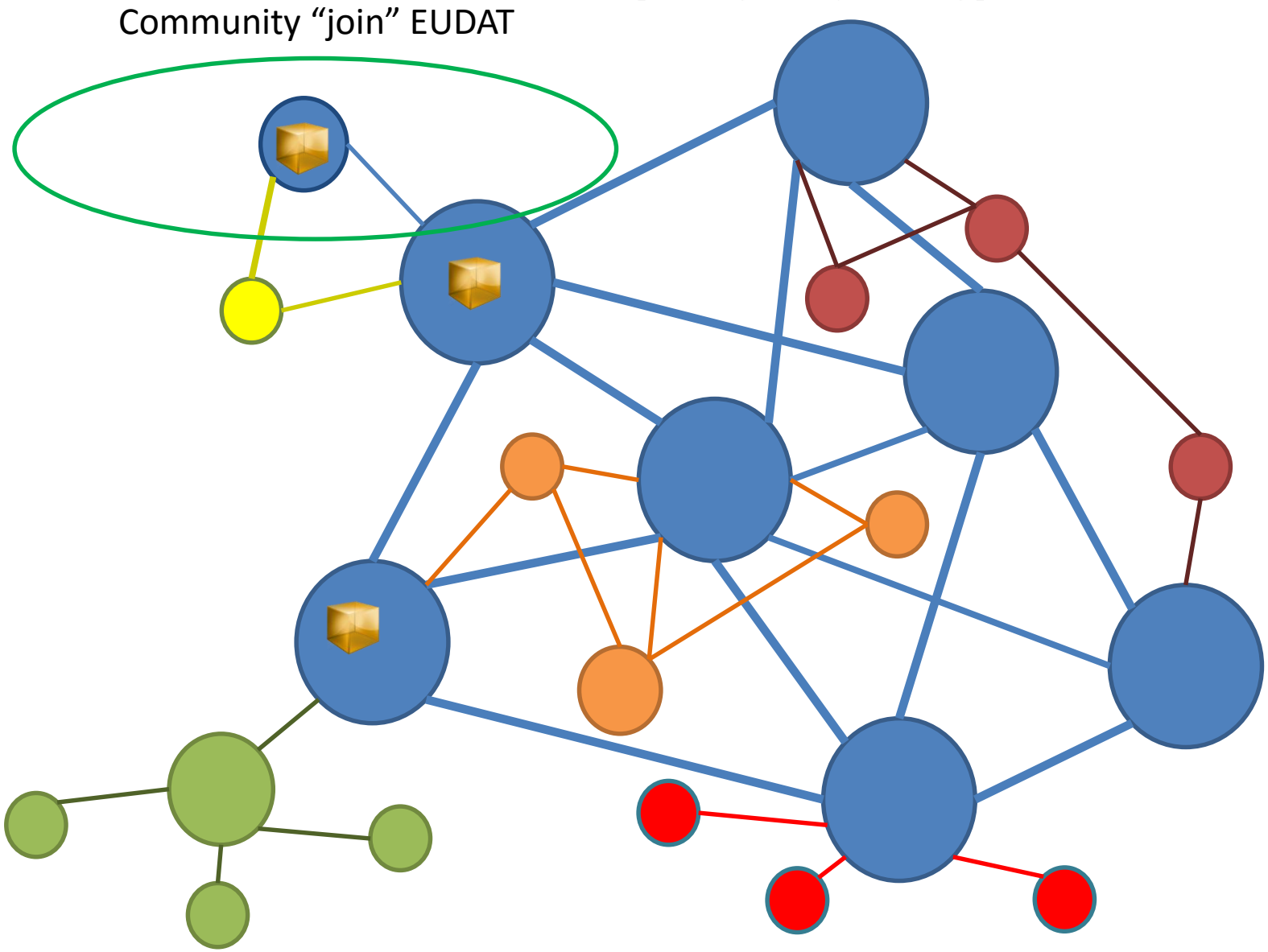
# EUDAT Collaborative Data Infrastructure
## (A general CDI architecture overview)



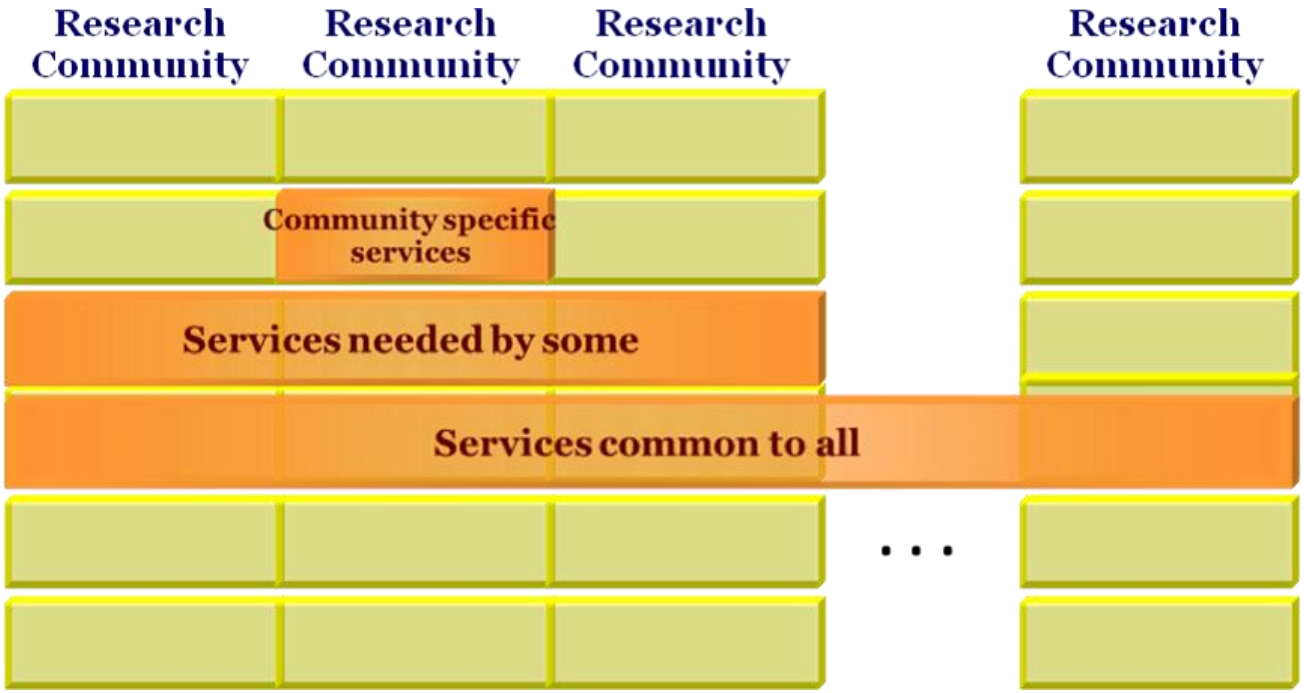
# EUDAT Collaborative Data Infrastructure
## (Using vs. joining)



# EUDAT Collaborative Data Infrastructure
## (Using vs. joining)



# If there are hundreds of Research Infrastructures, how many different data management systems can be sustained?



# B2 Service (modular) Suite

**B2DROP**
Sync and Exchange Research Data

**B2SHARE**
Store and Share Research Data

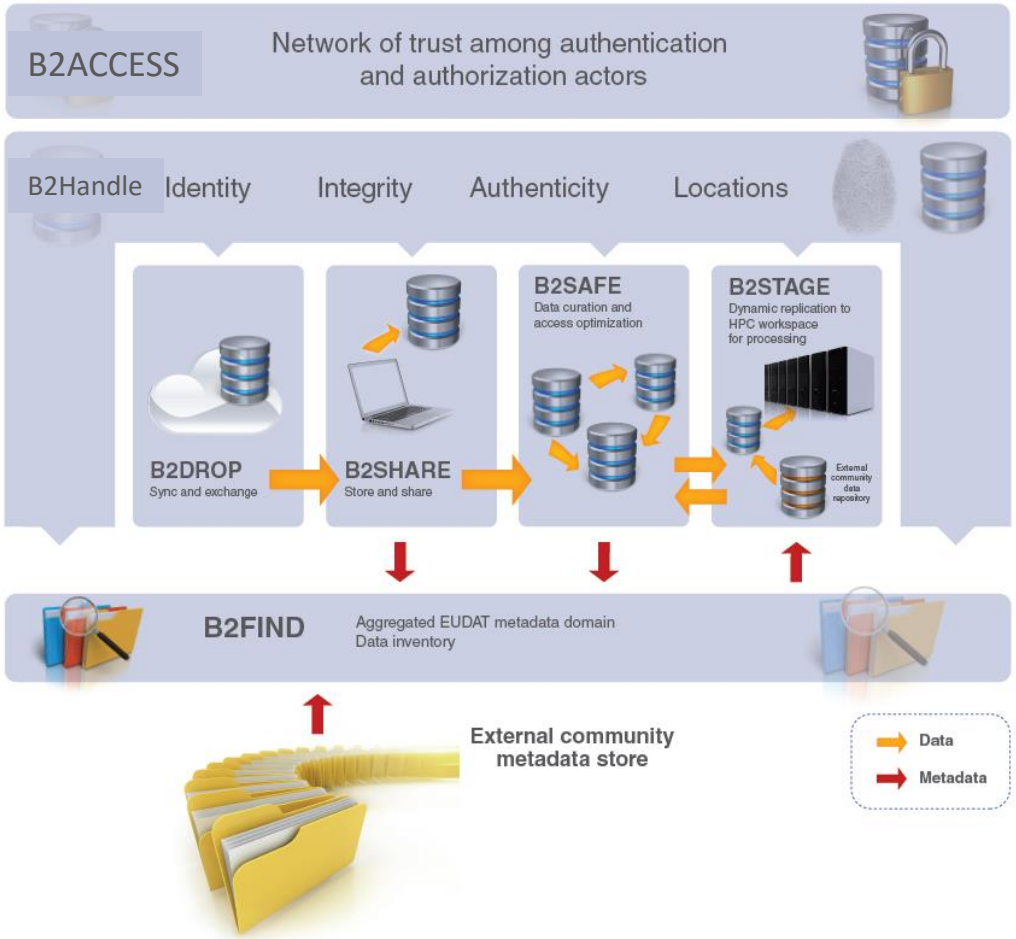
**B2SAFE**
Replicate Research Data Safely

**B2STAGE**
Get Data to Computation

**B2FIND**
Find Research Data

**B2HANDLE**
Register your Research Data

**B2ACCESS**
Identity & Authorisation

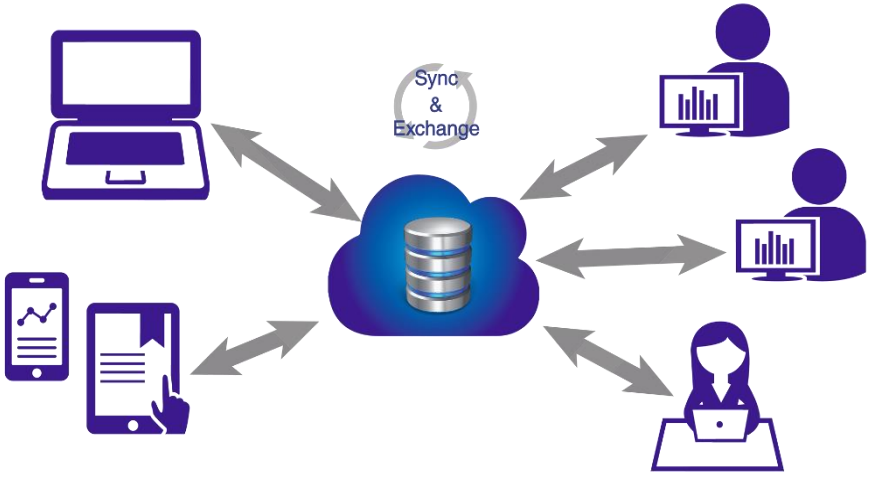


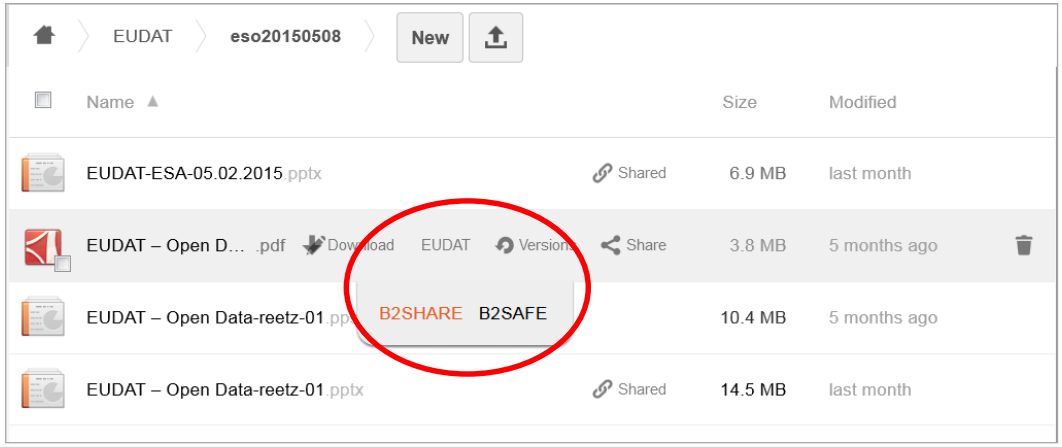
EUDAT

# B2DROP

Sync and Exchange Research Data

- **Who**
  - Citizens Scientists and small teams
- **What**
  - **Store and exchange** data
  - **Synchronize** multiple versions
  - Ensure **automatic desktop** synchronization
- **Why**
  - Ease of Use
  - Trusted European Service



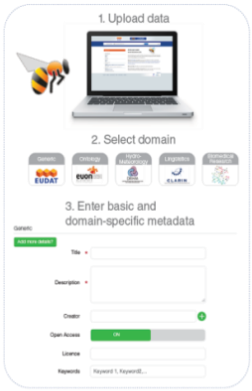

## EUDAT2020

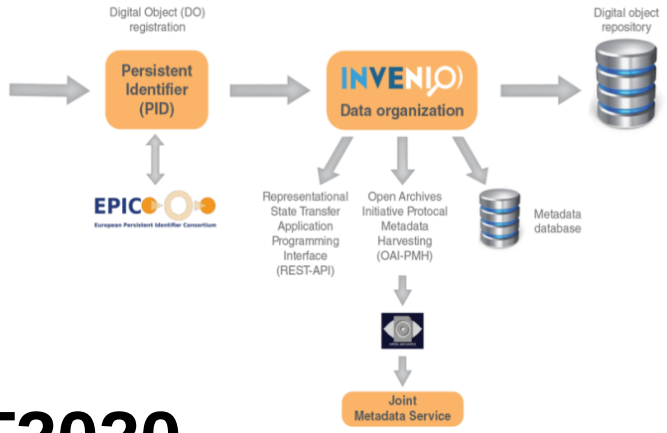
- Further **integration** with EUDAT CDI (e.g. B2SHARE)
- Integration with **B2ACCESS to enable access by many different Identity Providers**
- **Cloud Storage Federation**, collaboration with GEANT in OpenCloudMesh
- Assess B2DROP as **workspace** area to **computing facilities**

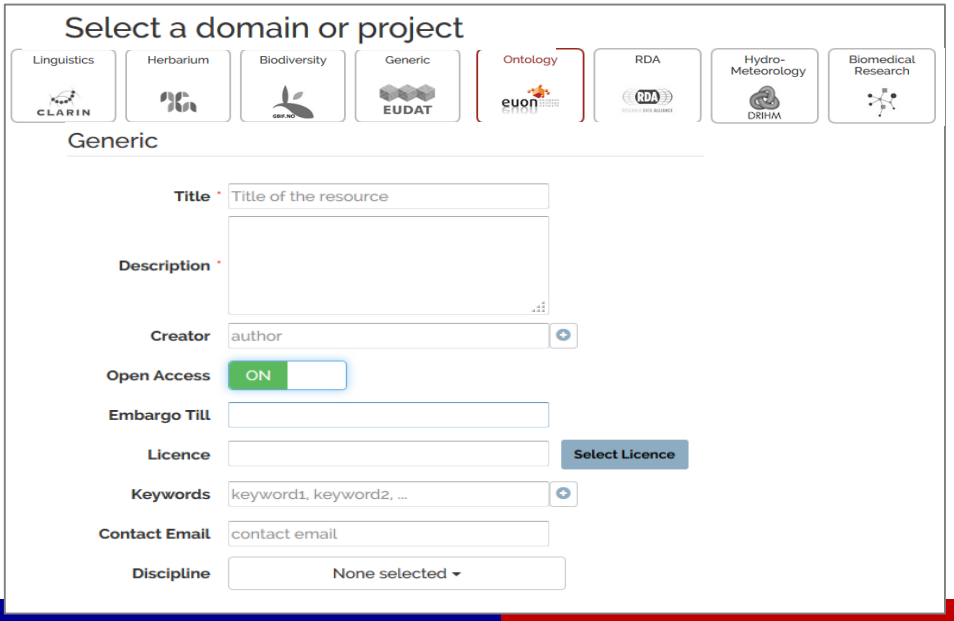
EUDAT

B2SHARE
Store and Share Research Data

- **Who**
  - Small to Medium Teams
- **What**
  - **Store** data (incl. software) and add domain meta data
  - **Share** registered research data worldwide
  - **Preserve** (small-scale) research data for long-term
- **Why**
  - Register Data for Publications
  - Make known to wider community



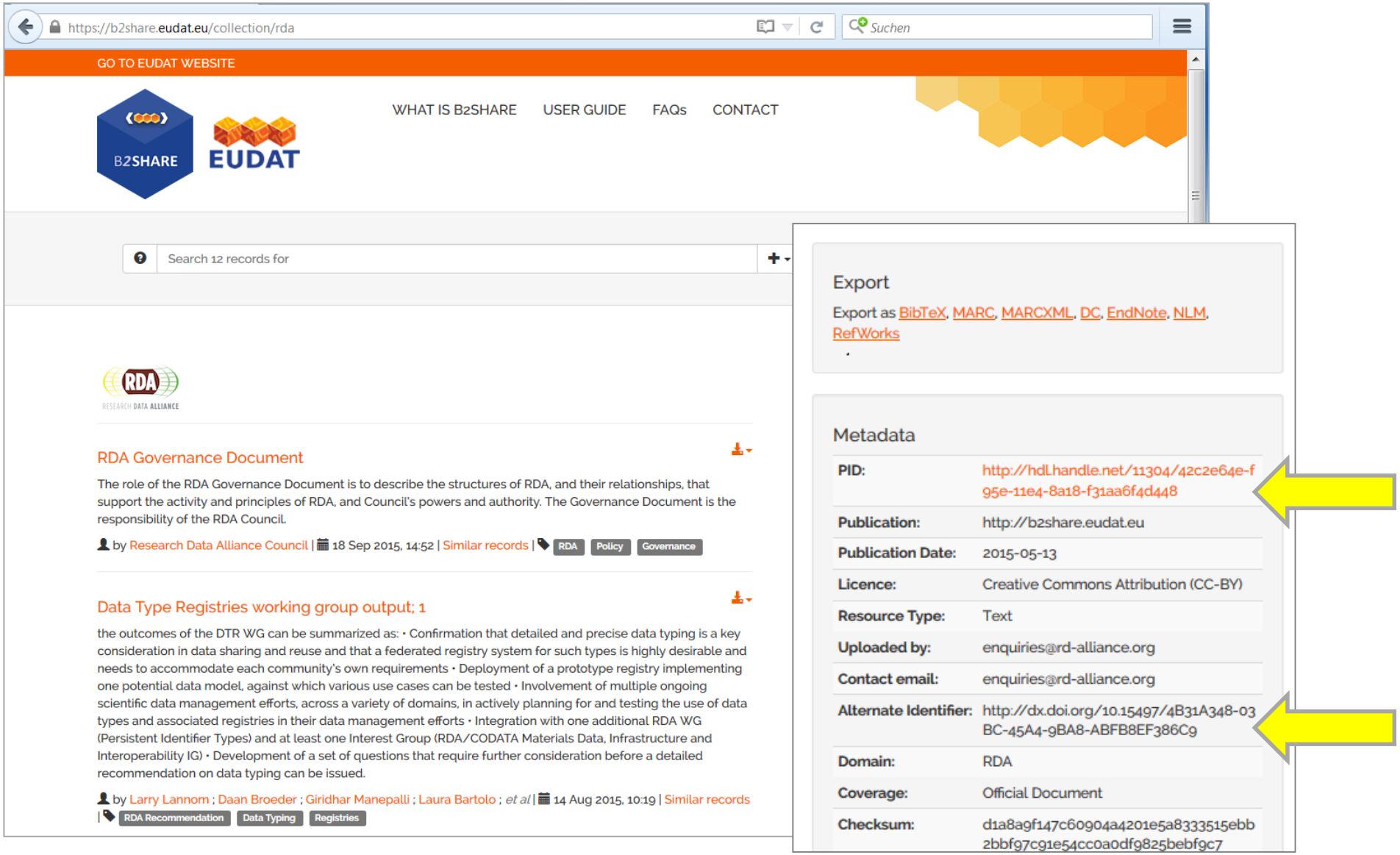




# EUDAT2020

- Further integration with EUDAT CDI (e.g. B2DROP, B2SAFE) WORK IN PROGRESS
- Integration with **B2ACCESS** (incl eduGAIN), focus on **authorization** WORK IN PROGRESS
- **Embargo** period
- **Editing of metadata**
- Data **versioning** and **annotation** WORK IN PROGRESS
- Extended HTTP Restful API interface WORK IN PROGRESS
- Easy installable **software package** WORK IN PROGRESS

# Collection of official RDA documents

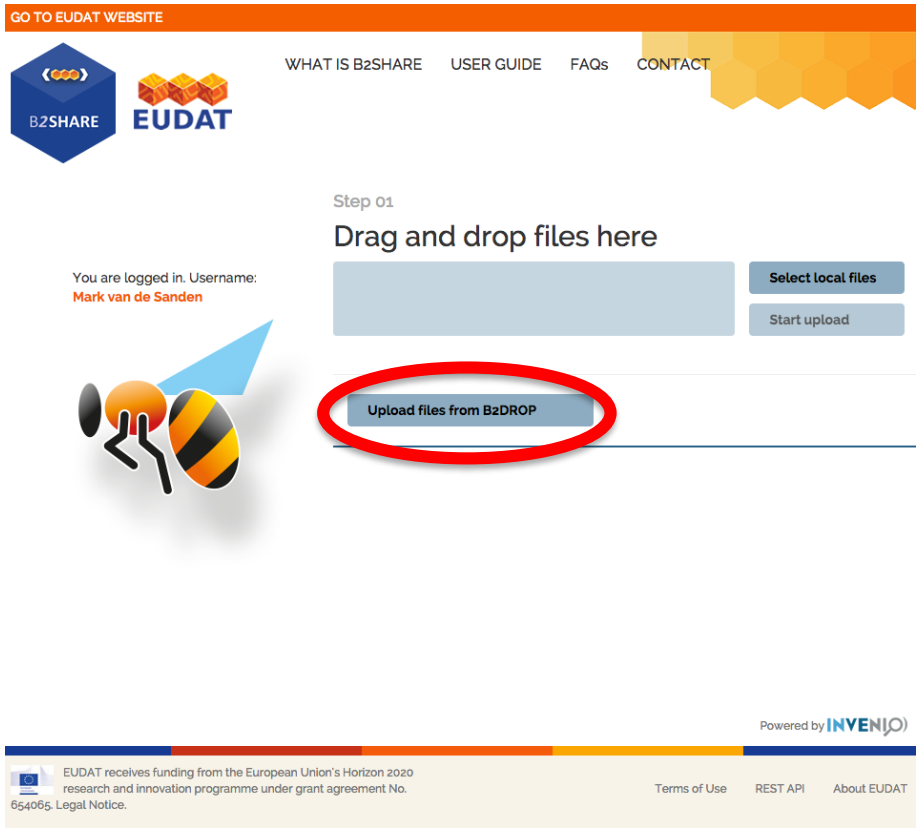


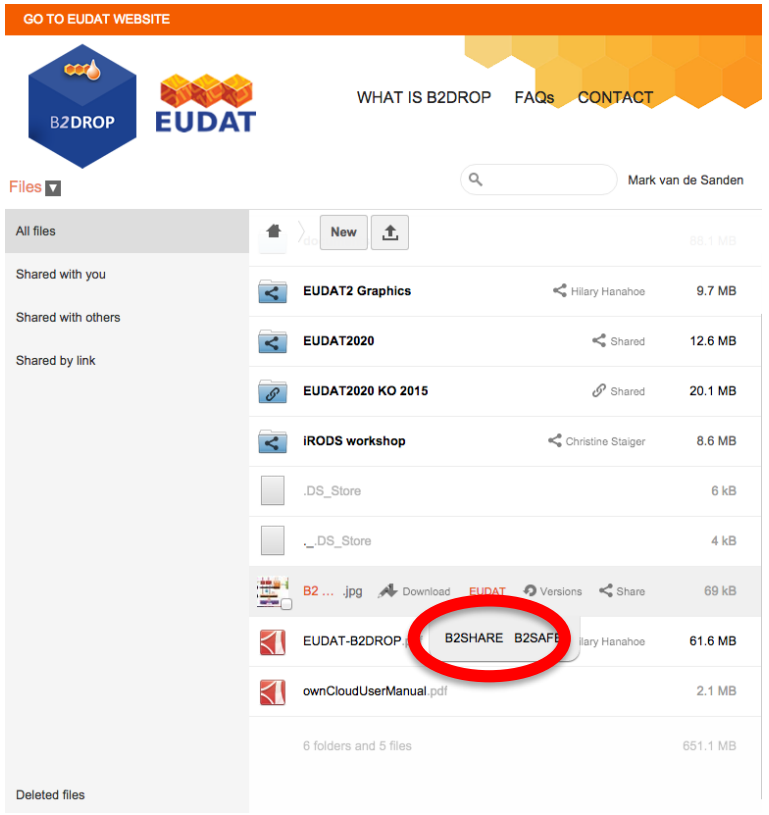
# Service Integration

B2SHARE
Store and Share Research Data

B2DROP
Sync and Exchange Research Data

Bidirectional Integration

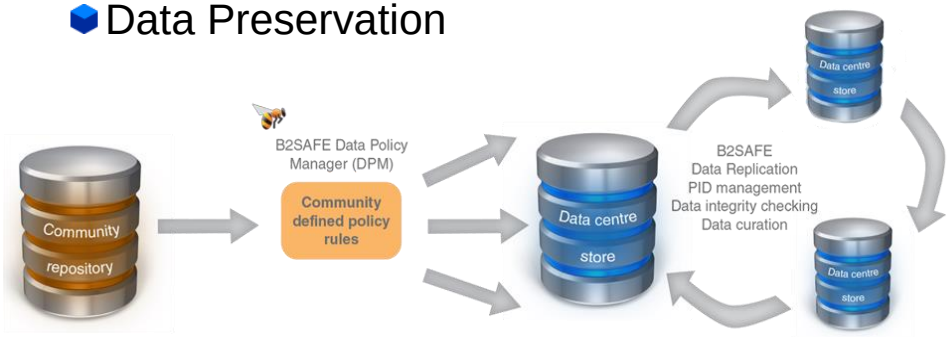


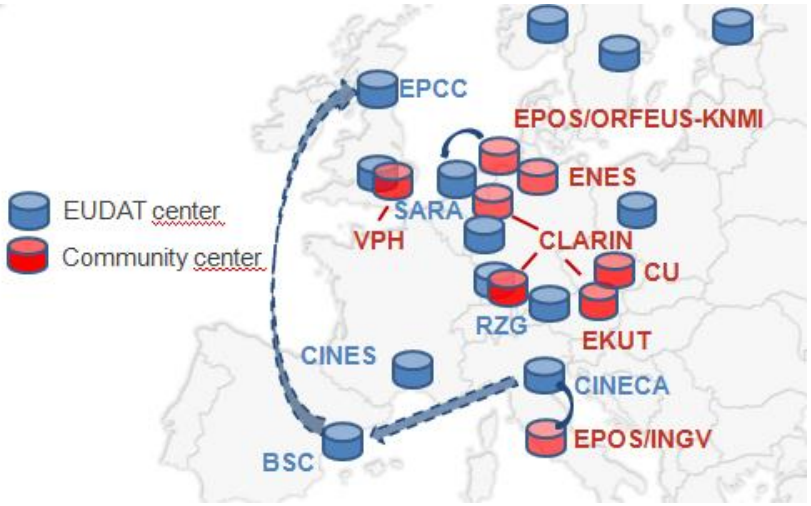



EUDAT

# B2SAFE
Replicate Research Data Safely

- **Who**
  - Community Data Managers
  - 'Sophisticated' Organisations
- **What**
  - Provide an abstraction layer which virtualizes large-scale data resources
  - Guard against data loss in long-term **archiving and preservation**
  - **Optimize access** for users from different regions
  - Bring data **closer to powerful computers**
- **Why**
  - Performance
  - Replication between trusted sites
  - Data Preservation




# EUDAT2020

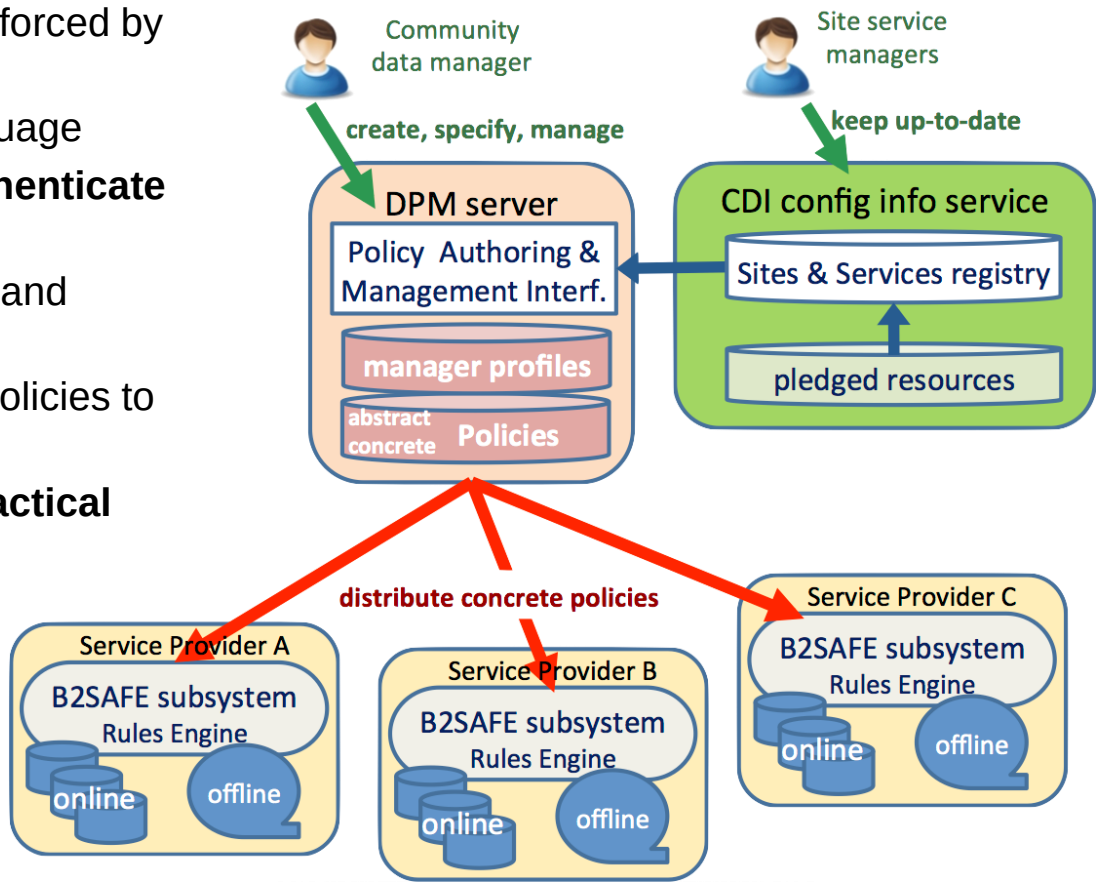
- Support **iRODS v4**
- Support **metadata**
- Optimize and **extend** policies to support **data curation** and **provenance**
- Further integration with **B2ACCESS**
- Support **authorization** on basis of **community access rules**
- Assess B2SAFE as **workspace** area to **computing facilities**

EUDAT

B2SAFE
Replicate Research Data Safely

# Data Policy Manager

- Data policies are **centrally managed**
- Policy rules are implemented and enforced by **site-local rule engines**
- Policies describe in an **abstract** language
- Community data managers must **authenticate** to provide **trust**
- Support policies for **data replication** and **integrity checking**
- **Central logging** for **auditable** data policies to **monitor** execution
- Active collaboration with the **RDA Practical Policy WG**



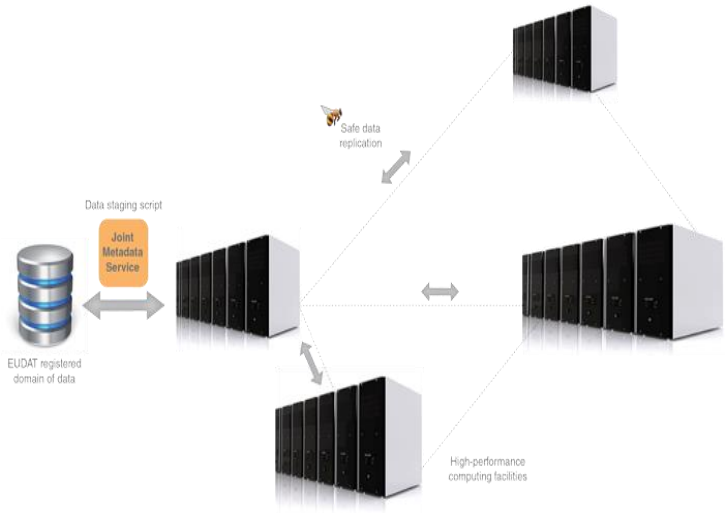

# EUDAT2020

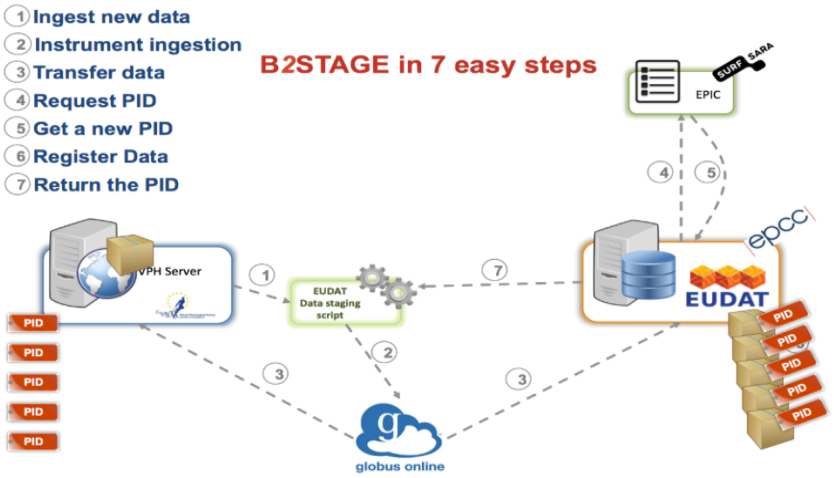
- **Handover** to operations WORK IN PROGRESS
- **Extend** number of policies supported
- Focus on **data curation** and **provenance** policies
- Integrate with **B2ACCESS**

EUDAT

# B2STAGE
Get Data to Computation

- **Who**
  - Users and Communities with Significant Computational Needs
- **What**
  - **Transfer** large data collections from EUDAT storages to **external HPC facilities for processing**
  - **Copy large data sets, ingesting them** onto EUDAT storage resources
- **Why**
  - Integration/Collaboration with PRACE
  - Simplify Data Transfer



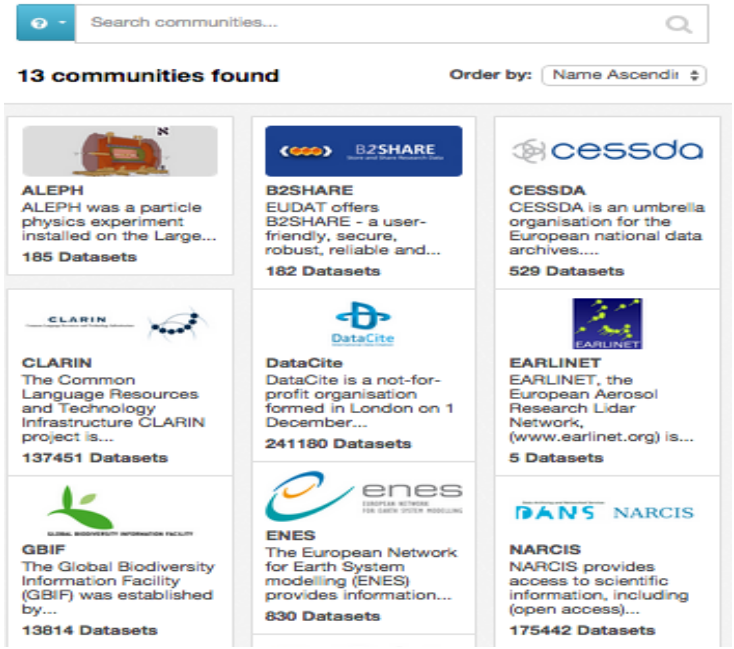

# EUDAT2020

- Further develop **HTTP** to a **mature interface** and extend functionality to **metadata**
- **Native support PIDs** within **GridFTP** transfers
- Extend EUDAT **client API** library to other **B2 services** (e.g. B2SHARE, B2FIND, PID)
- Further integration with B2ACCESS

EUDAT

B2FIND
Find Research Data

- **Who**
  - Anyone
- **What**
  - **Find** collections of scientific data quickly and easily, irrespective of their origin, discipline or community
  - Get quick **overviews** of available data
  - Browse through collections using **standardized facets**
- **Why**
  - Unique collection
  - Ease of Searching



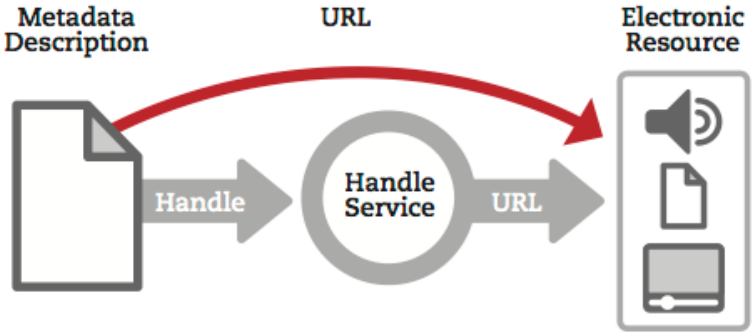
# EUDAT2020

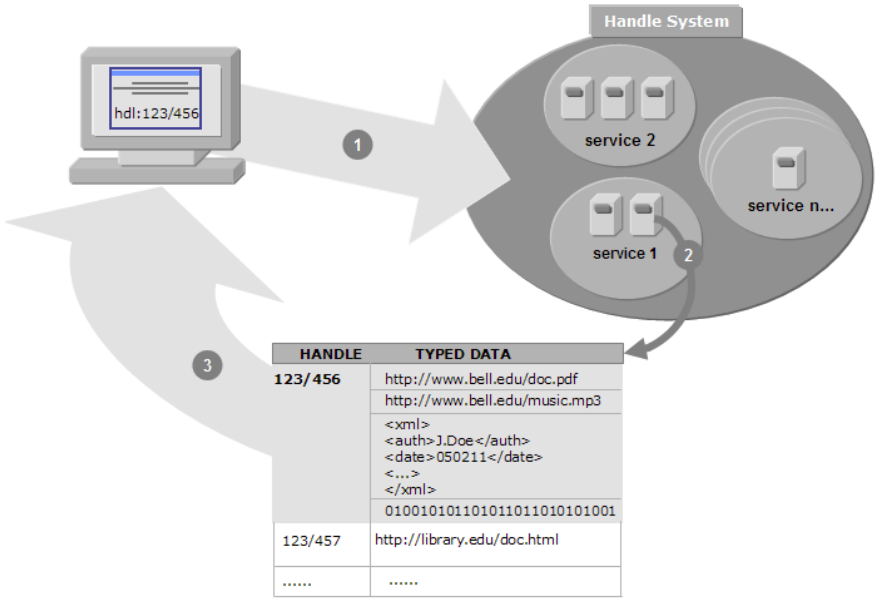
- Harvesting of **metadata** stored in **B2SAFE**
- **Community customizations**
- **Annotation** of datasets
- Further assess **RDF** and **Linked Data**
- Further assess **scalability** and **performance**

EUDAT

B2HANDLE
Register your Research Data

- **Who**
  - Groups or Communities who want to make their data citable
- What
  - Follows **policies** to register data and make it **long term refer**- and **citable**
  - Reliability through mutual **PID mirroring**
  - Provides **abstraction layer** between a globally **unique persistent identifier** and **physical location** of data objects
  - M**achine readable** via **HTTP RESTful API**
- **Why**
  - Simple integration
  - Technology Agnostic



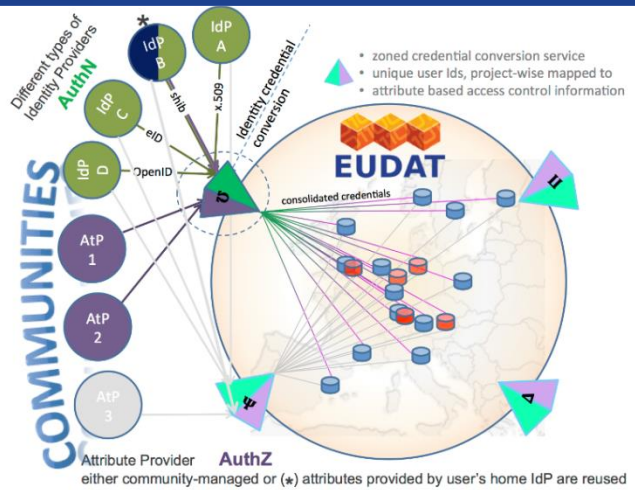

# Development plan

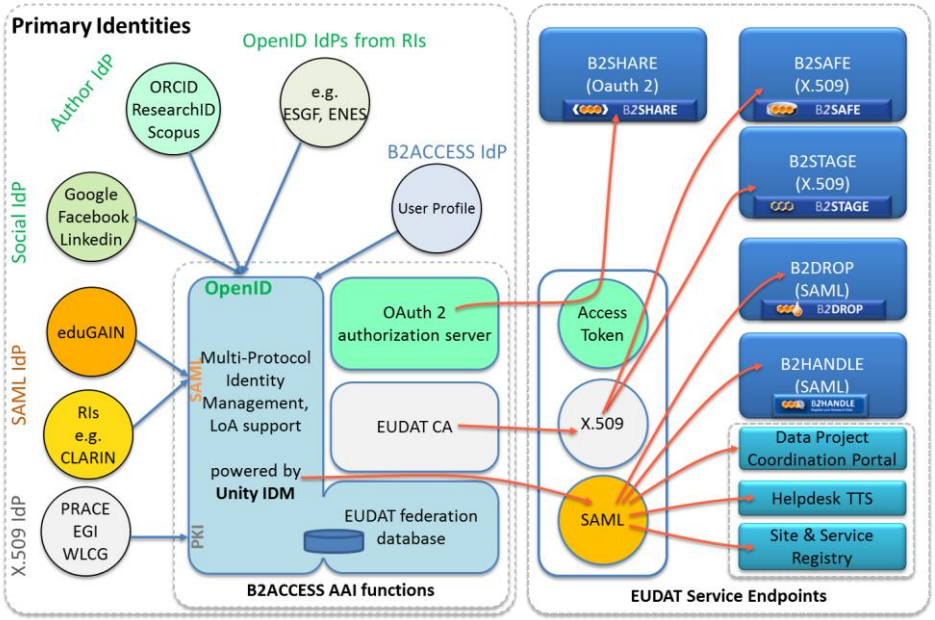
- Develop the policies for the B2HANDLE service (e.g. PID namespace mngmt) WORK IN PROGRESS
- Migrate service from Handle v7 to v8 WORK IN PROGRESS
- Define PID Information Types for data, metadata, collection records WORK IN PROGRESS
- Integrate with Data Type Registry service
- Consolidate B2HANDLE API library with EUDAT API library WORK IN PROGRESS

EUDAT

# B2ACCESS
Identity & Authorisation

- **Who**
  - Anyone wanting to use the B2 Services
- **What**
  - Complies with **community ownerships** and **access rights**, basis of trust
  - Credential **conversion approach** (e.g. SAML, OpenID, X.509, Username/password)
  - Identity provider for **citizen scientists**
- **Why**
  - Use your own ID in federated environment




# EUDAT2020

- **Integration** with operational and all **B2 services**
  **B2SHARE** ✓ **B2DROP** WORK IN PROGRESS **B2STAGE** WORK IN PROGRESS
  **B2SAFE** WORK IN PROGRESS **B2HANDLE, DPM, CREG** ✓**, TTS,**
- **Integration** with **community IdP domains** and **portal environments**
- Enabling access via eduGAIN ✓ social IDs ✓
  enabling access via ORCID WORK IN PROGRESS CLARIN IdPs
- Focus on **authorization**
- Collaborate on **cross e-infrastructure access** WORK IN PROGRESS
  (e.g. PRACE, EGI)
- Extend **European collaboration** via **AARC** WORK IN PROGRESS
  (e.g. Geant, Terena)

B2ACCESS EUDAT



Register a new account

Username:

Password:

Authenticate

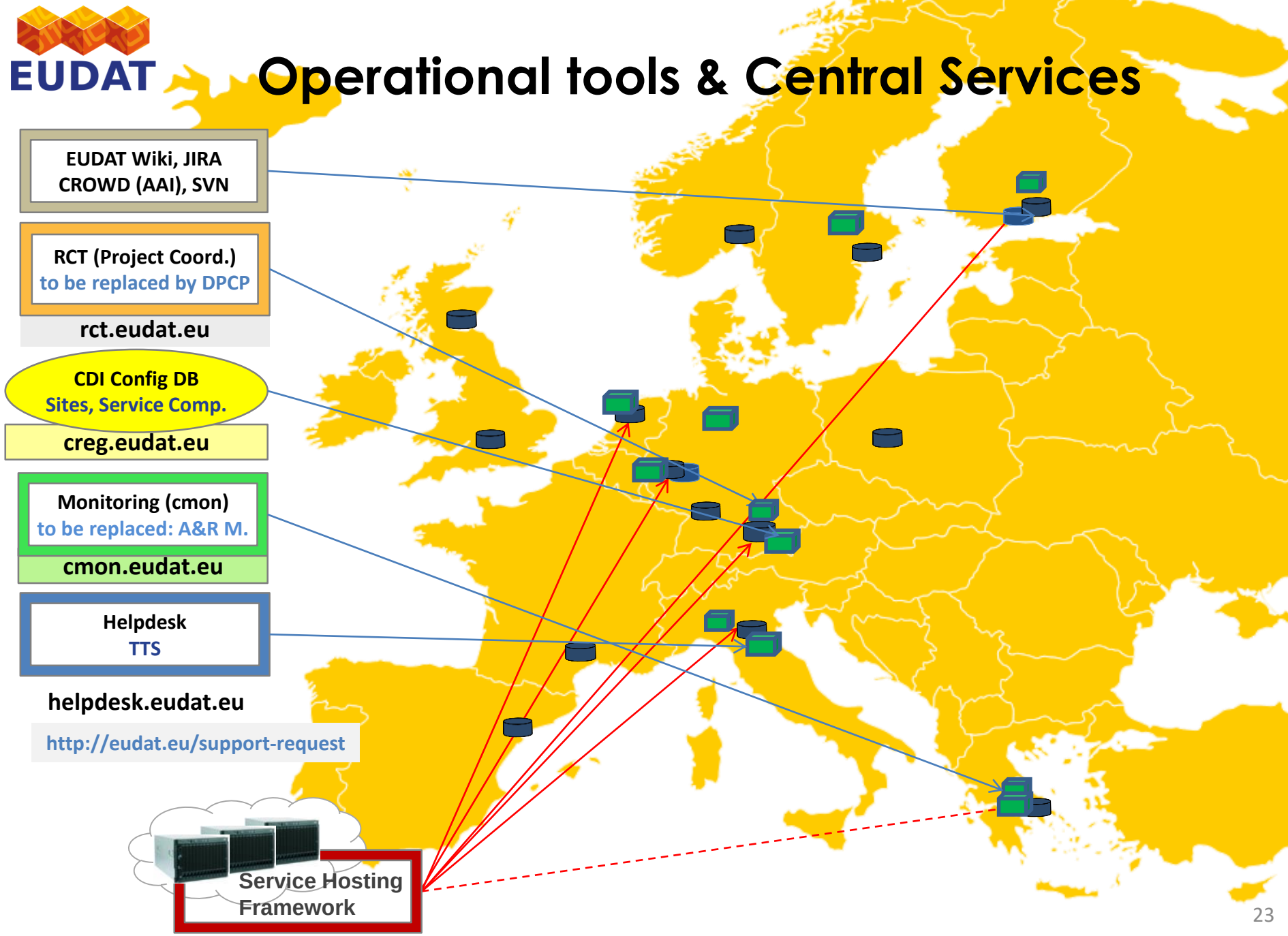
## Select your preferred authentication option

Search:

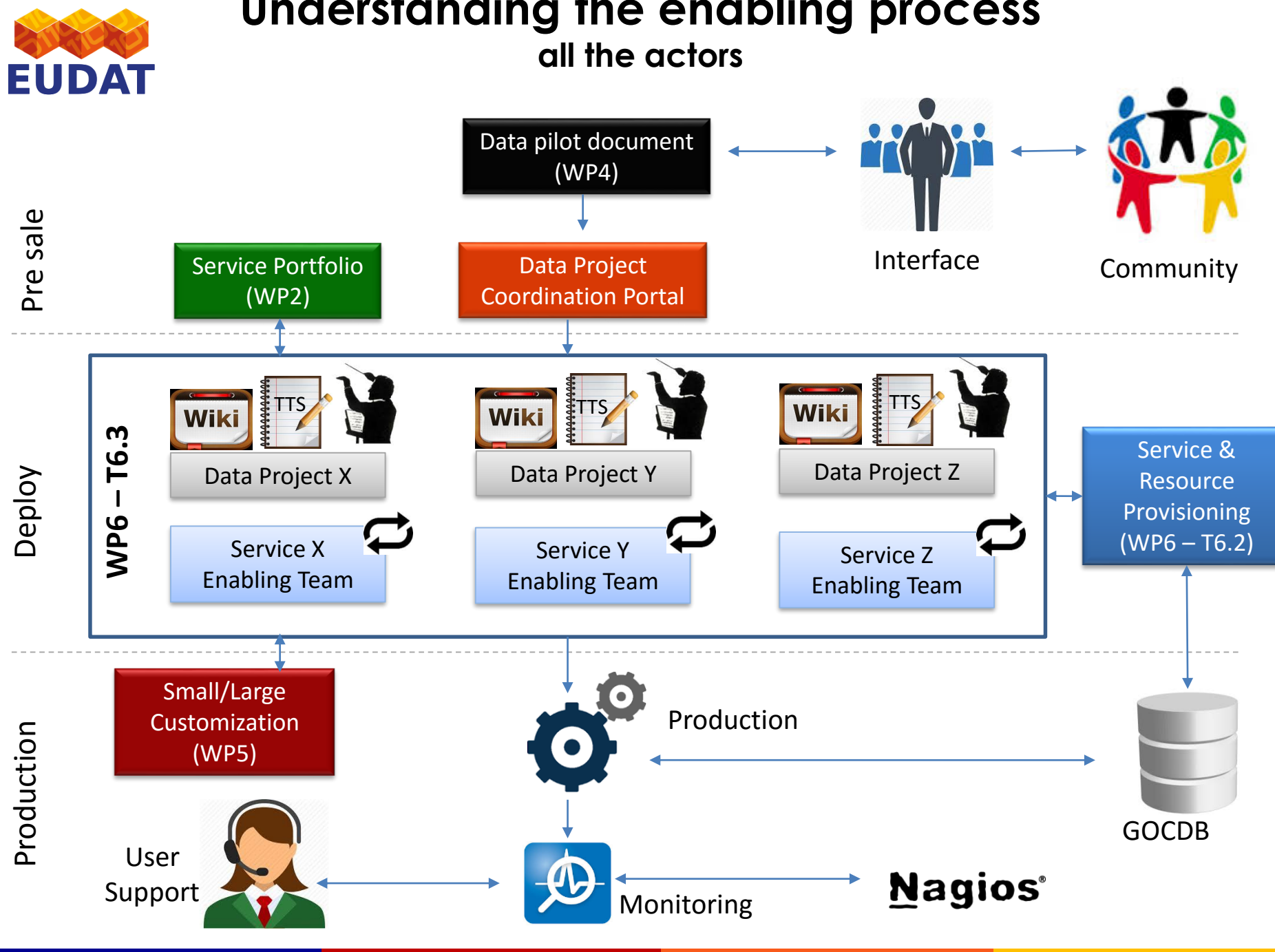
| Log in with your B2ACCESS ID | Log in with your social ID | Log in with your organisational ID |
|---|---|---|
| Password | Google Account | Czech Technical University in Prague |
| | | Czech Technical University in Prague, Faculty of Electrical Engineeri... |
| | | DARIAH |

EUDAT receives funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 654065. Legal Notice

# Operational tools & Central Services



# Understanding the enabling process
## all the actors



# Understanding the enabling
## Deploy actors

Deploy

**WP6 – T6.3**

Data Project X

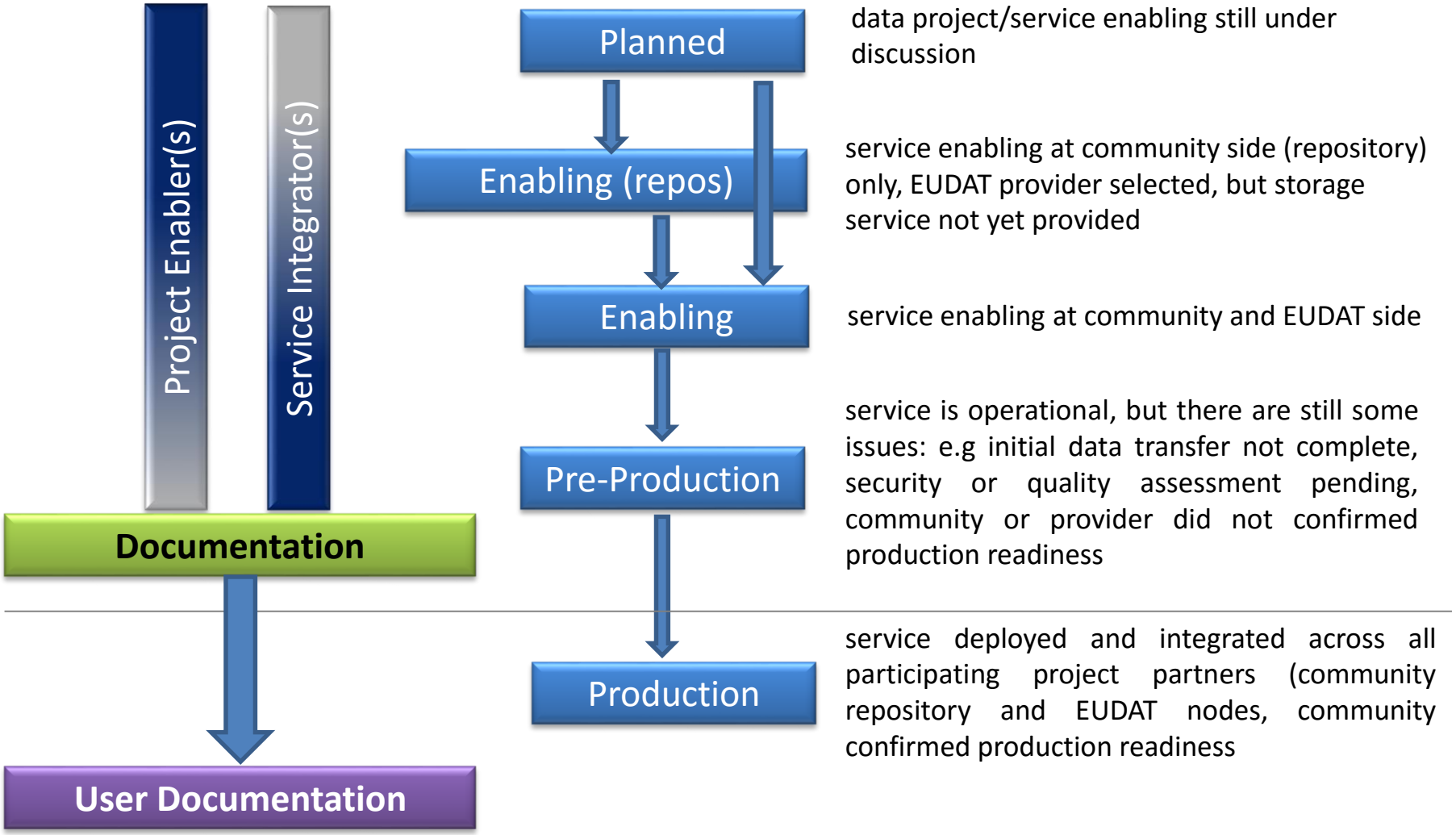
**Project Enabler**

**Service Integrator**

Service X Enabling Team

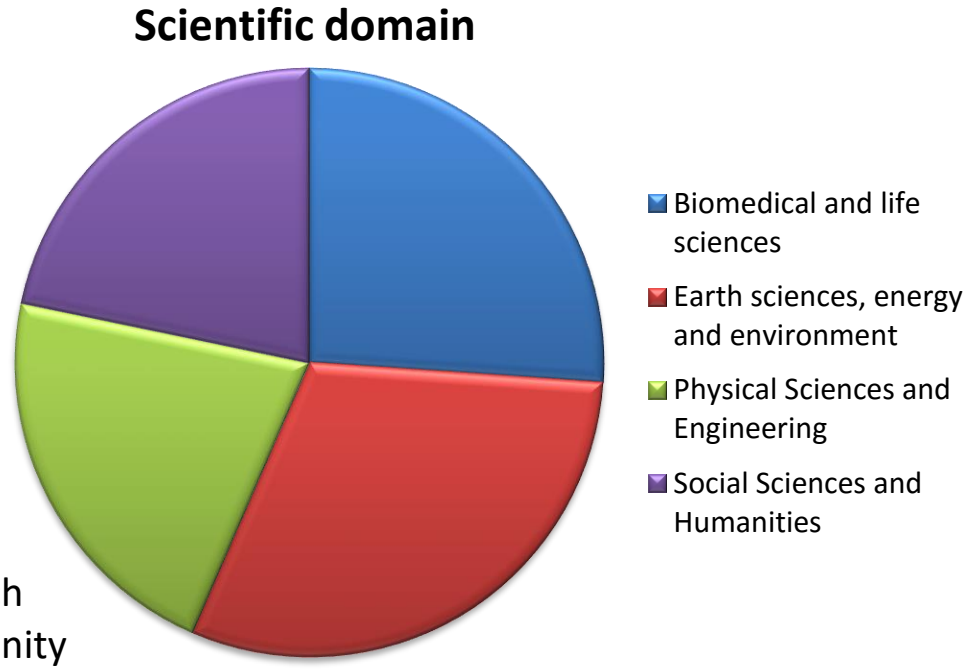
Service integration into community

# Understanding the enabling
## Project Lifecycle and relationship with Project Enablers and Service Integrators

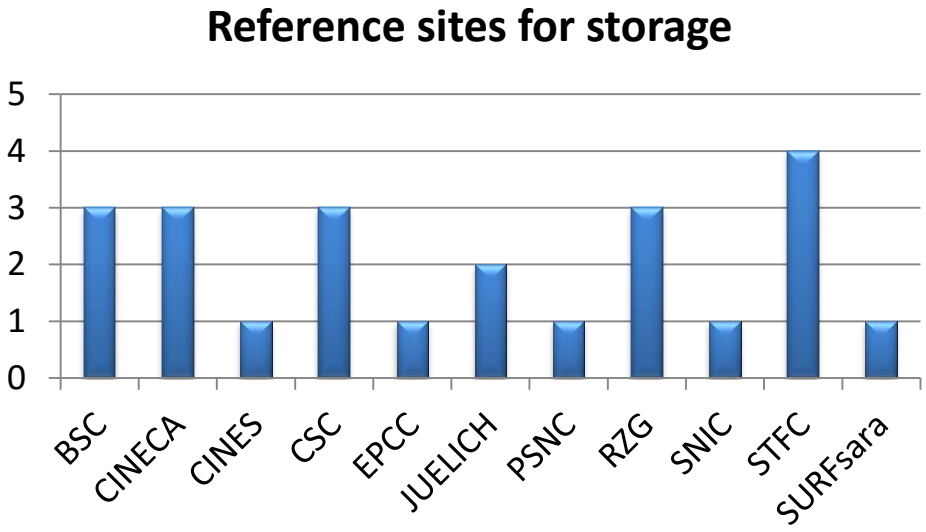


# Data pilots overview

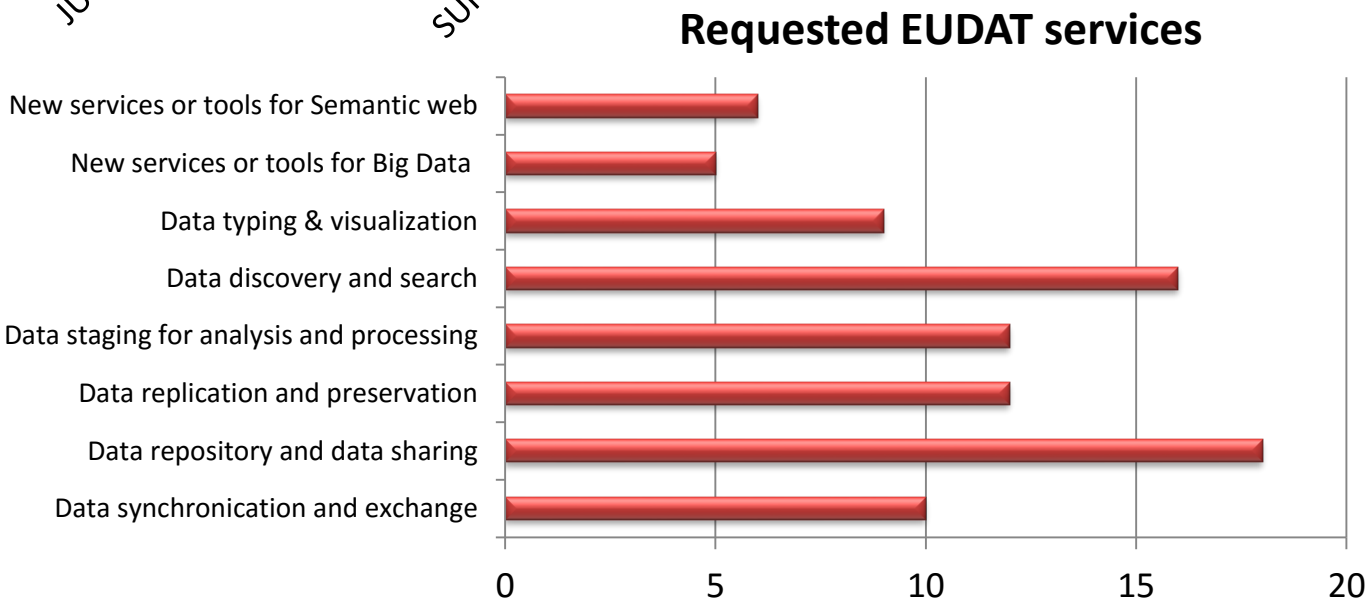
- 23 data pilots selected for enabling in EUDAT2020



# Data pilots overview




Total storage request 1220-4300 TB



# Questions…