EOSC-hub Data Platforms for data processing and solutions for publishing and archiving scientific data (part 1)

NARGES ZARRABI

SARA RAMEZANI

EOSC-HUB Week

12th April 2019

Objective of the session

- **Part 1:** Show how EUDAT services can be used for managing active research data and for preserving final research data (i.e. data archiving and publishing). We also demonstrate how these services operate and integrate with each other comply with the FAIR principles.
- **Part 2:** Demonstrate how end-users can perform data analysis on large volume of datasets, and produce reusable results following the FAIR principles.

Audience: This training track is relevant for researchers, IT support people, and service providers who operate services for Open Science.

Outline

- Data management requirements of research communities (10’)
- Overview of B2Services for data management (30’)
  - B2DROP, B2SAFE, B2SHARE, B2STAGE, B2HANDLE, B2FIND, B2ACCESS, B2NOTE…
  - Integration between B2Services
- Example data pipelines and workflows (Live demo) (40’) –
  - Use Cases:
    - CompBioMed: Safe data replication with B2SAFE
    - SeaDataCoud
  - Hands On:
    - Data sharing and publishing workflow with B2DROP and B2SHARE
    - Data publication with B2SHARE (API demo)
- Q&A (10’)

Data Requirements of research communities

- **More efficient data access and sharing**
  - *Intensive data-sharing*
  - *Restricted data-sharing*
- **Preserving research data**
  - *Storage, backup and archiving large data, synchronizing data over distributed places*
  - *data provenance*
- **Accessible research Data**
  - *Making data accessible to research communities, PIDs*
  - *Publishing data with domain specific metadata*
  - *Linking published data to processed and raw data*
- **Findable research data**
  - *A major challenges scientific communities is to discover data from research data collections and repositories*

Data Life Cycle

EUDAT Collaborative Data Infrastructure

CREATING DATA: designing, planning consent, collection and management, capturing and creating metadata

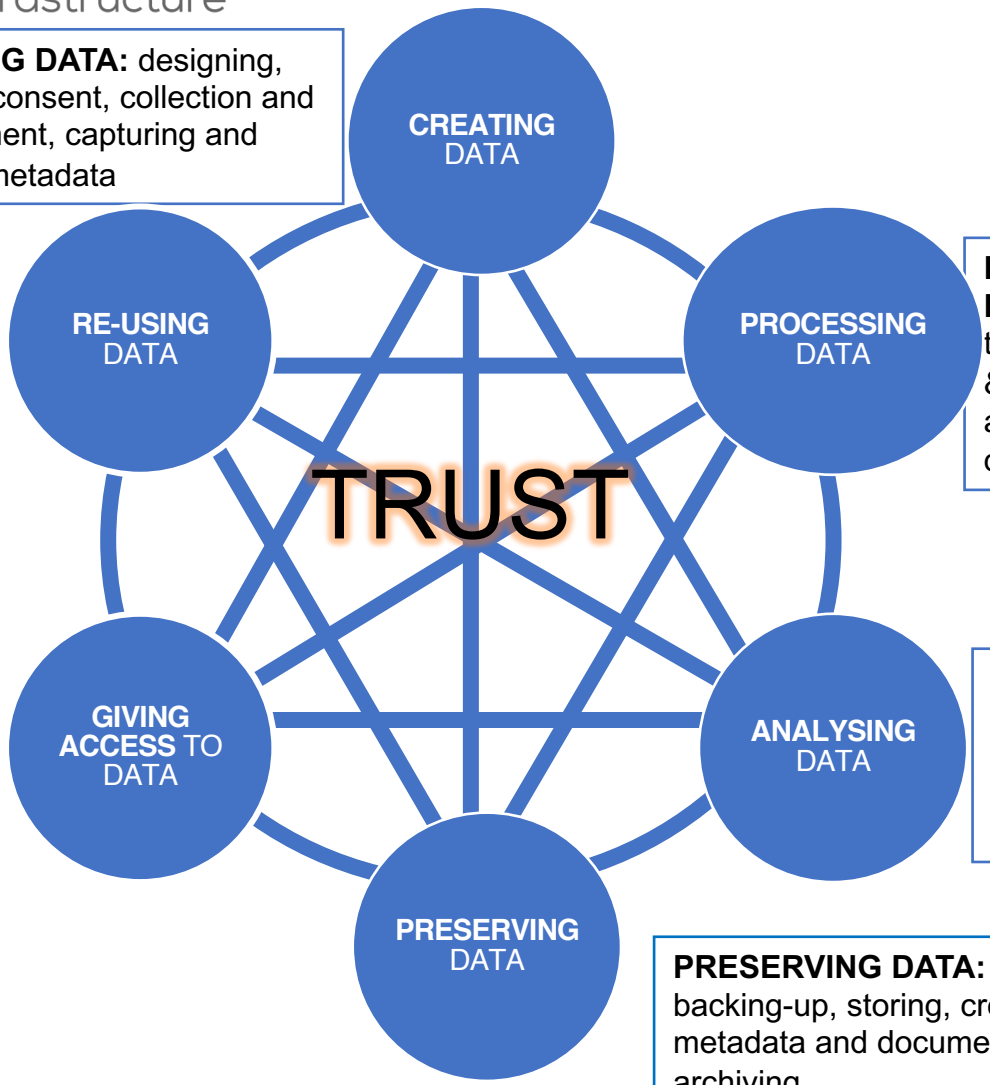
PROCESSING DATA: entering, transcribing, checking & validating, anonymizing and describing

ANALYSING DATA: interpreting, deriving, producing outputs & publishing, preparing for sharing

PRESERVING DATA: migrating, backing-up, storing, creating metadata and documentation, archiving

GIVING ACCESS TO DATA: distributing, sharing, controlling access, promoting

RE-USING DATA: for follow-ups, new research, research reviews, scrutinizing, teaching & learning



Ref: UK Data Archive: http://www.data-archive.ac.uk/create-manage/life-cycle

What is… FAIR ?

Findable:

F1. (meta)data are assigned a globally unique and persistent identifier;

F2. data are described with rich metadata;

F3. metadata clearly and explicitly include the identifier of the data it describes;

F4. (meta)data are registered or indexed in a searchable resource;

Accessible:

A1. (meta)data are retrievable by their identifier using a standardized communications protocol;

A1.1 the protocol is open, free, and universally implementable;

A1.2. the protocol allows for an authentication and authorization procedure, where necessary;

A2. metadata are accessible, even when the data are no longer available;

Interoperable:

I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.

I2. (meta)data use vocabularies that follow FAIR principles;

I3. (meta)data include qualified references to other (meta)data;

Reusable:

R1. meta(data) are richly described with a plurality of accurate and relevant attributes;

R1.1. (meta)data are released with a clear and accessible data usage license;

R1.2. (meta)data are associated with detailed provenance;

R1.3. (meta)data meet domain-relevant community standards;

doi: 10.1038/sdata.2016.18

EUDAT B2Service Suite

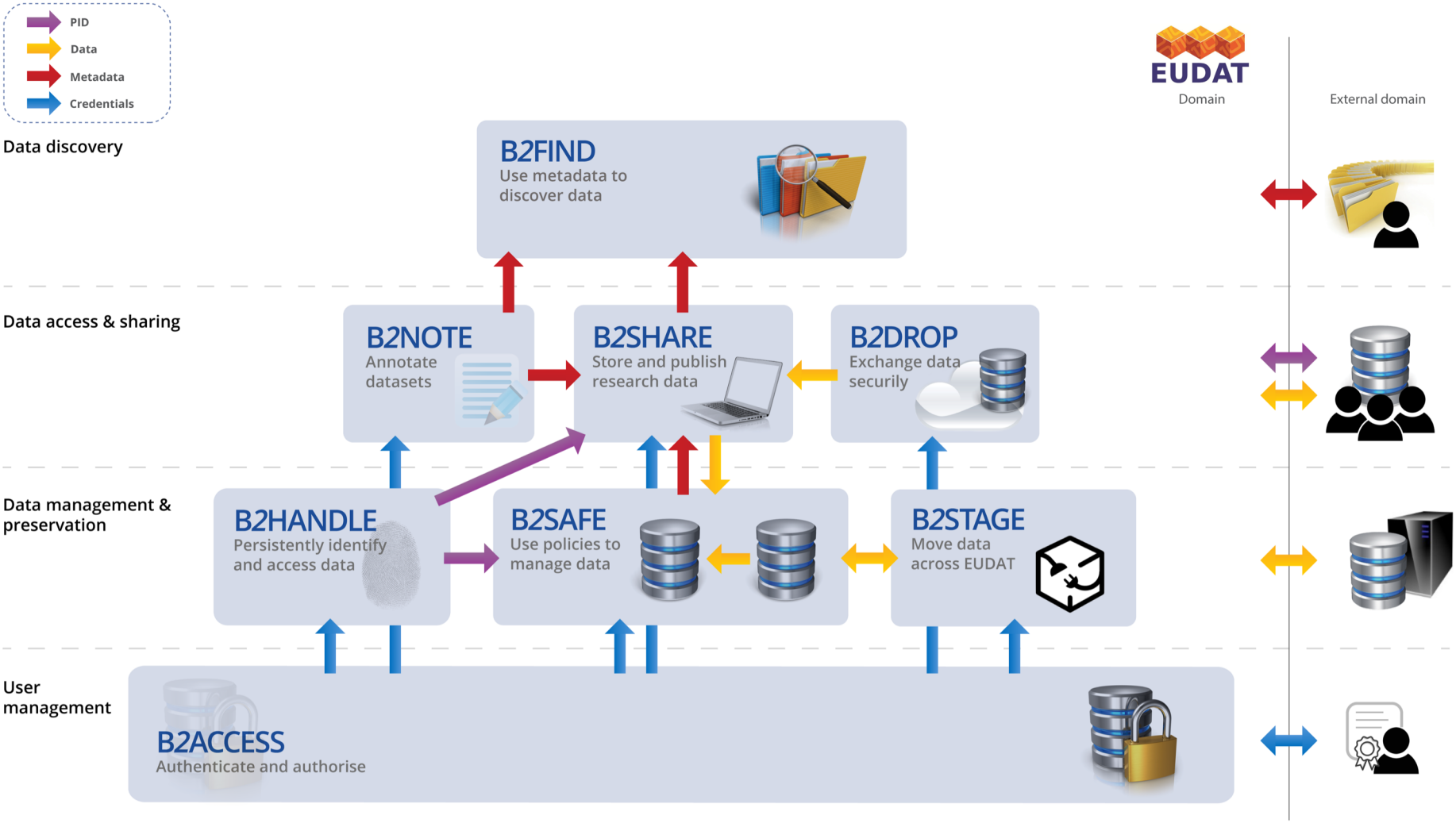
- **EUDAT B2Service Suite**
  - **B2DROP**
  - **B2HANDLE**
  - **B2SAFE**
  - **B2STAGE**
  - **B2SHARE**
  - **B2NOTE**
  - **B2FIND**

- **How EUDAT services link to data lifecycle**

- **How EUDAT services support the FAIR principles**
  - **Helping scientists to generate FAIR data**

EUDAT B2 Service Suite
(An overview)

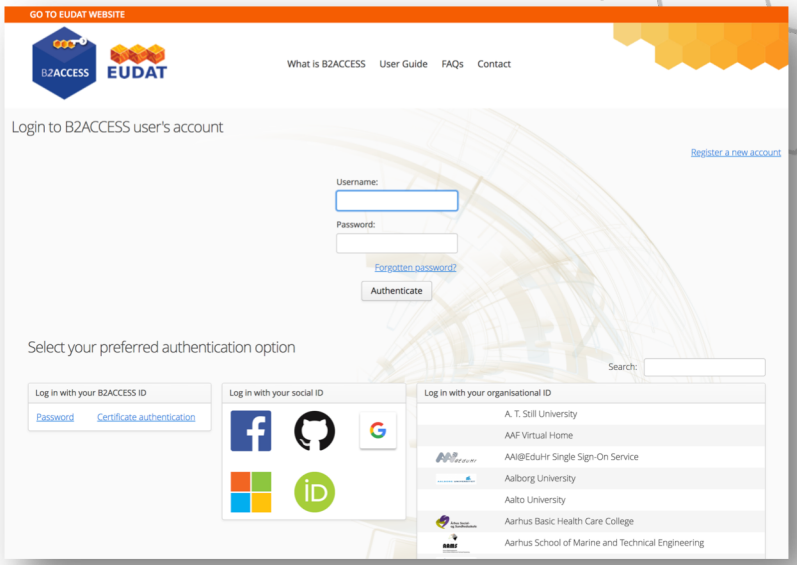
EUDAT B2 Services Diagram

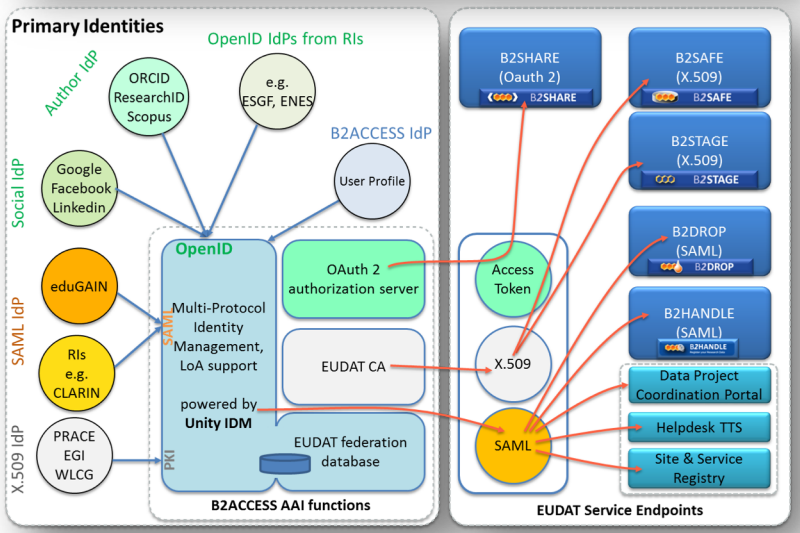


EUDAT Collaborative Data Infrastructure

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



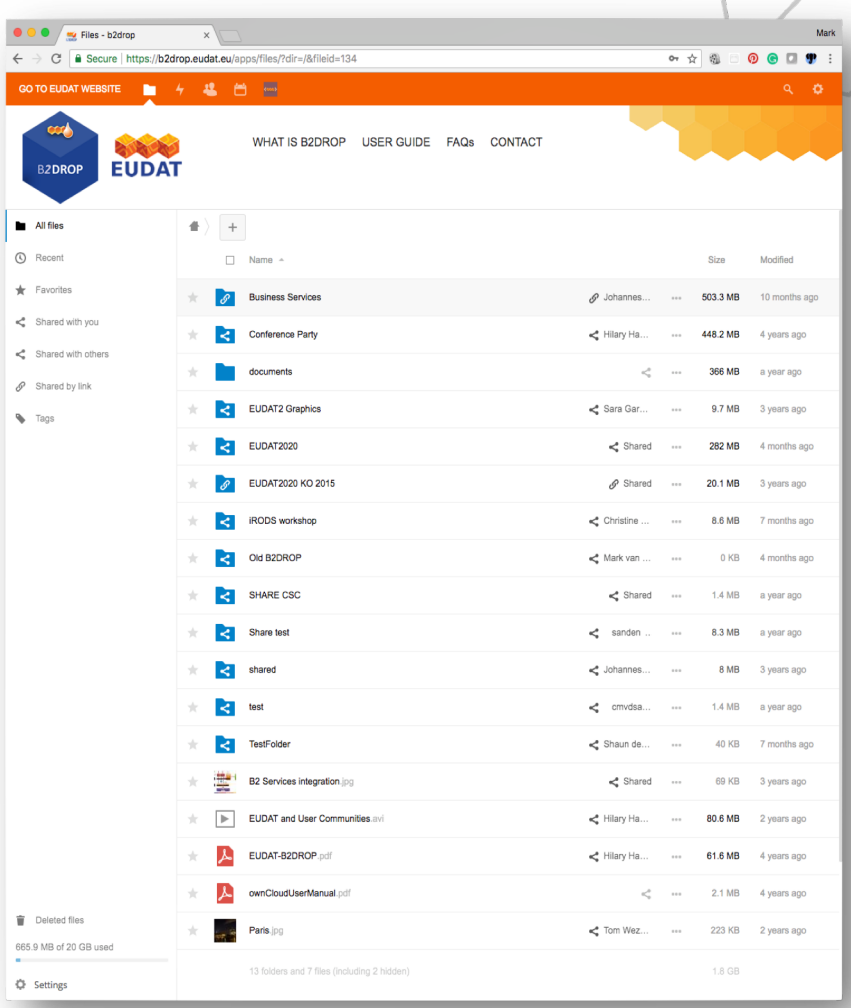


https://b2access.eudat.eu/

EUDAT Collaborative Data Infrastructure

B2DROP
Sync and Exchange Research Data

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

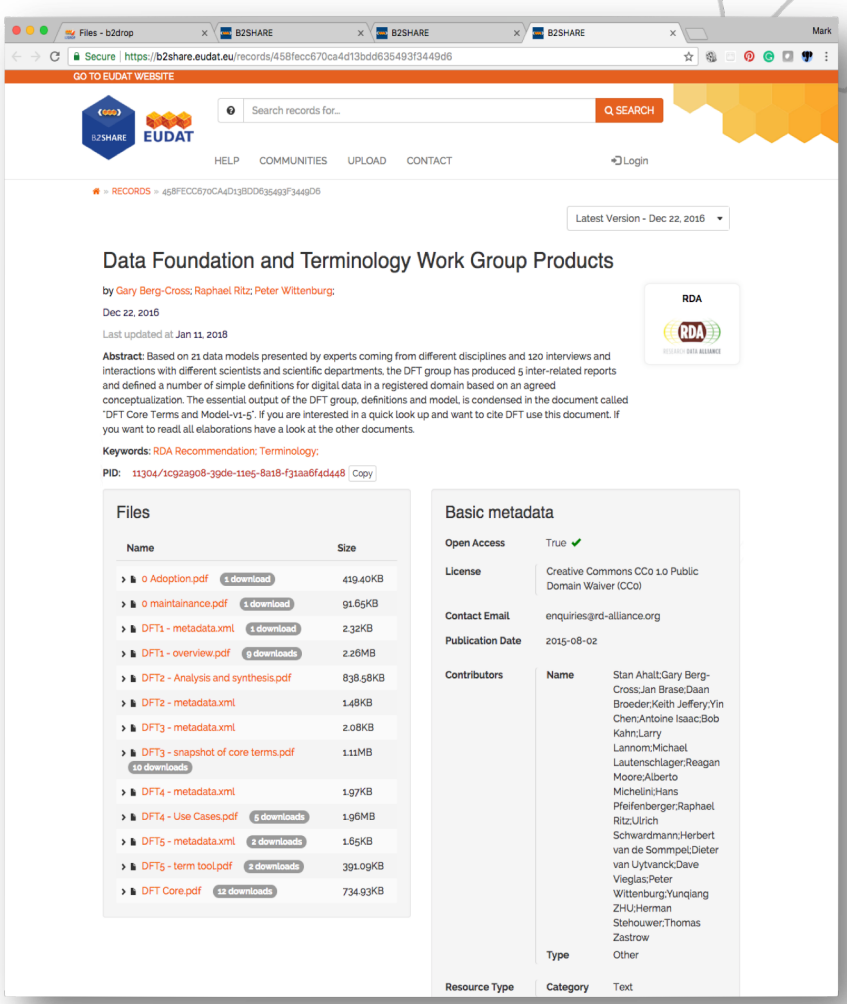


https://b2drop.eudat.eu/

EUDAT Collaborative Data Infrastructure

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 (FAIR)
  - Make known to wider community

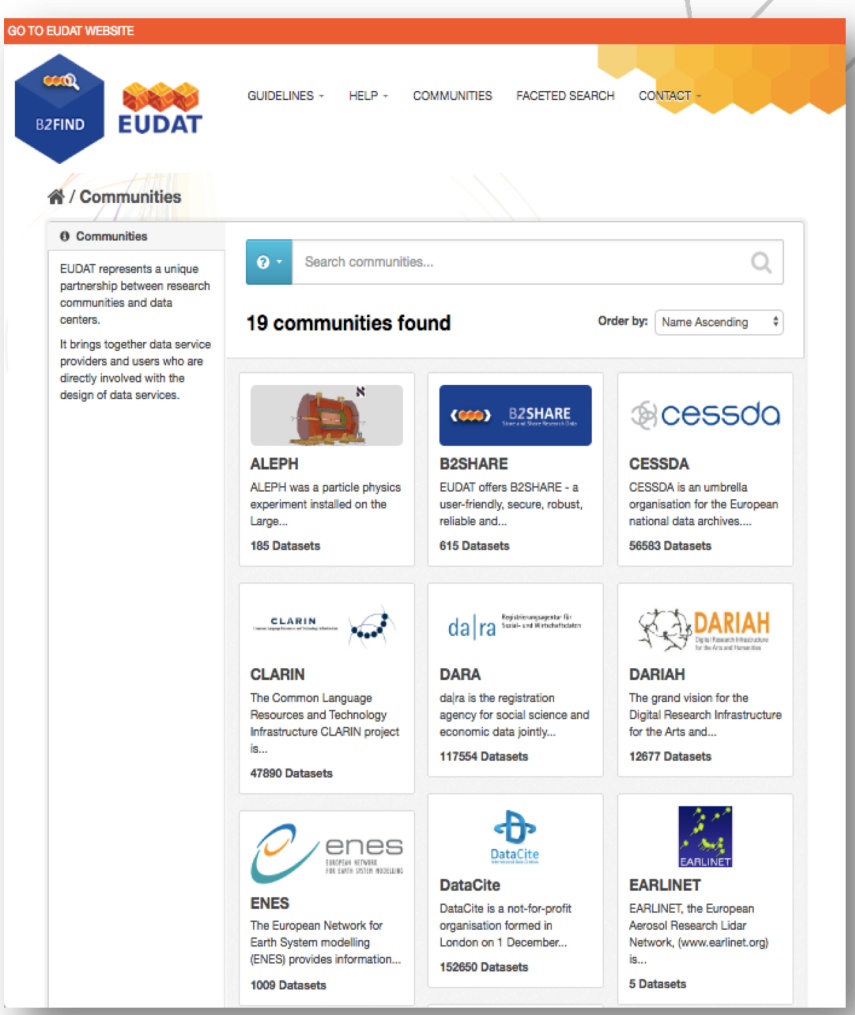


https://b2share.eudat.eu/

EUDAT Collaborative Data Infrastructure

B2FIND
Find Research Data

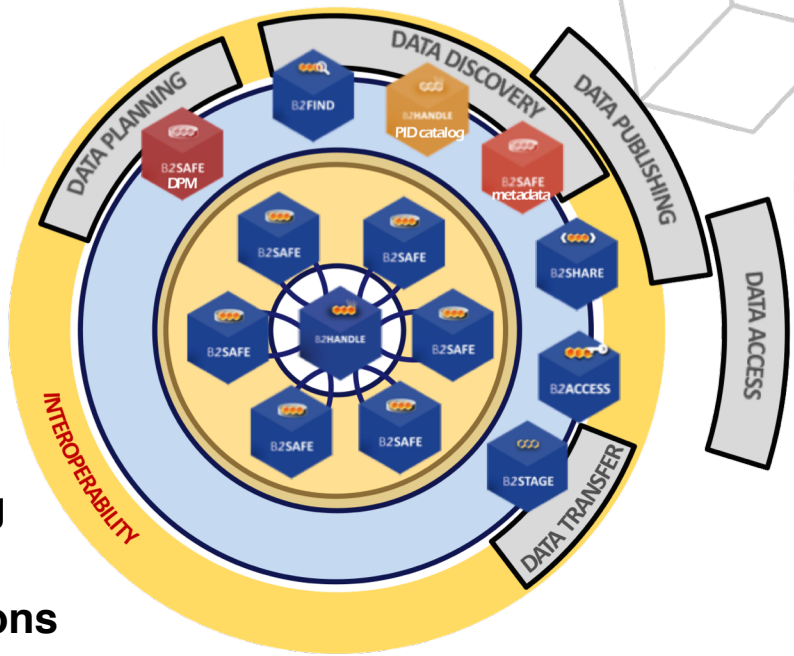
- **Who**
  - Anyone
- **What**
  - **Find** collections of scientific data quickly and easily, irrespective of their origin, discipline or community
  - Get quick **overviews** of available data
  - Browse through collections using **standardized facets**
- **Why**
  - Unique collection
  - Ease of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



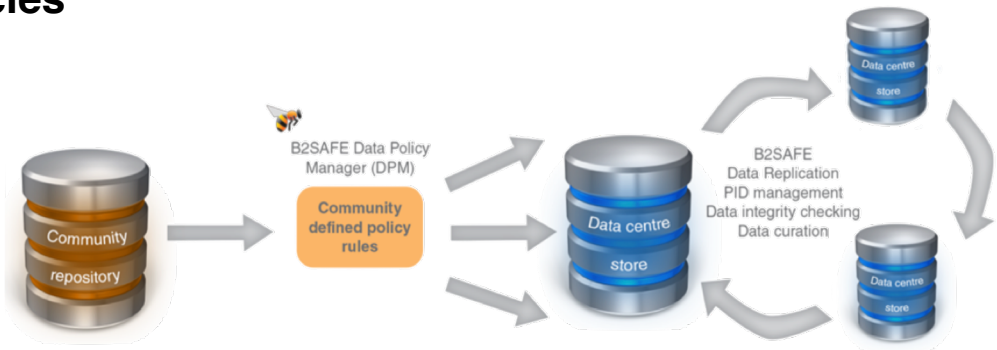
http://b2find.eudat.eu/

EUDAT Collaborative Data Infrastructure

B2SAFE
Replicate Research Data Safely



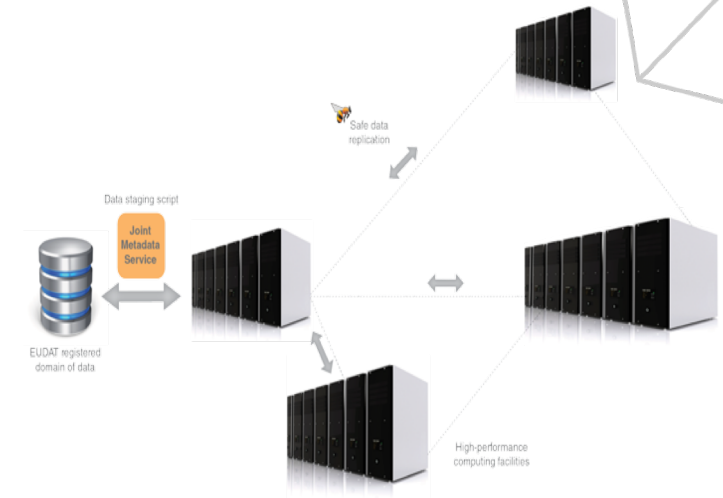
- **Who**
  - Community Data Managers
  - ‘Sophisticated’ Organizations
- **What**
  - Provide an abstraction layer which virtualizes large-scale data resources
  - Guard against data loss in long-term **archiving and preservation**
  - **Optimize access** for users from different **regions** and to **computing** resources
  - Data management on basis of **policies**
- **Why**
  - Performance
  - Replication between trusted sites
  - Data Preservation

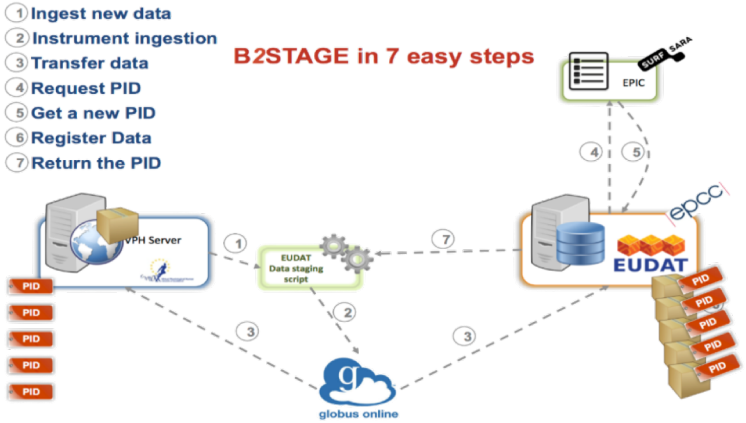


EUDAT Collaborative Data Infrastructure

B2STAGE
Get Data to Computation

- **Who**
  - Users and Communities who want to interact with EUDAT CDI services
- **What**
  - Provide a **common access** layer to B2 services
  - **Copy large data sets, ingesting them** onto EUDAT data services
  - Enables **data transfer** for large data collections from EUDAT storages to **external HPC facilities for processing**
- **Why**
  - Support data transfers between PRACE and EGI
  - Simplify data transfers





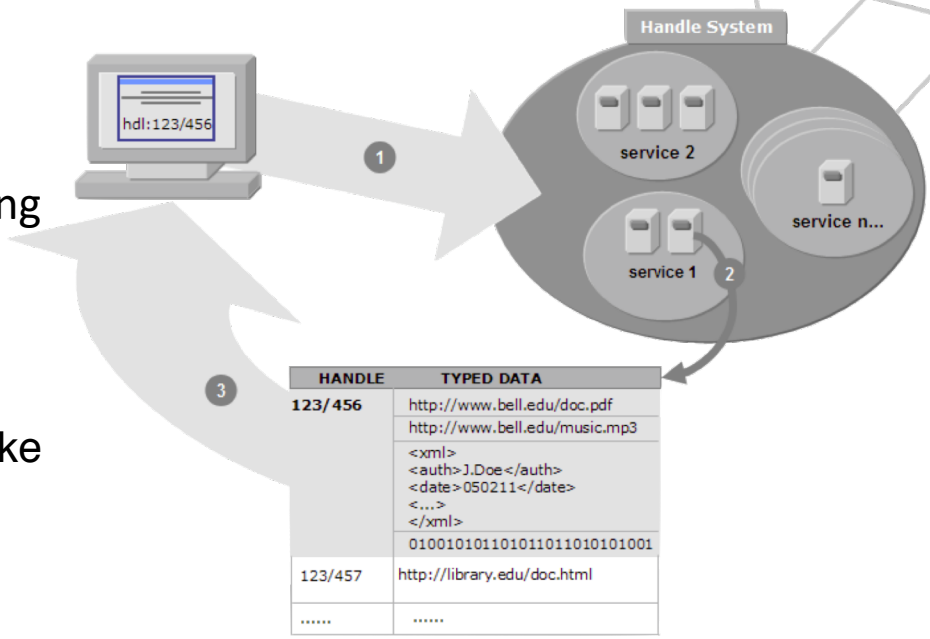
http://petstore.swagger.io/?url=https://b2stage.cineca.it/api/specs&docExpansion=none - /

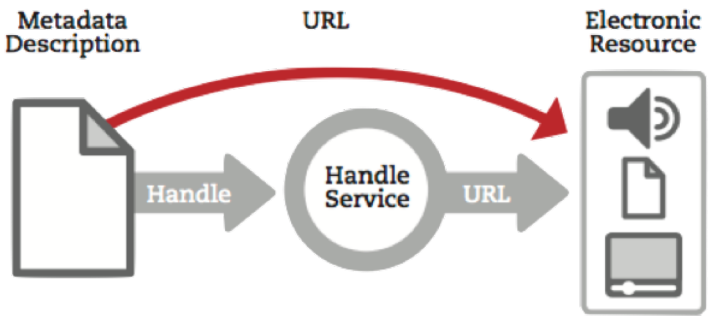
EUDAT Collaborative Data Infrastructure

B2HANDLE

Register your Research Data

- **Who**
  - Groups or communities who want to make their data **referenceable**, improving **data management tasks**
- **What**
  - Follows **policies** to register data and make it **long term referenceable**
  - Reliability through mutual **PID mirroring**
  - Provides **abstraction layer** between a globally **unique persistent identifier** and **physical location** of data objects
  - PIDs **global resolvable**
- **Why**
  - Simple integration
  - Technology Agnostic

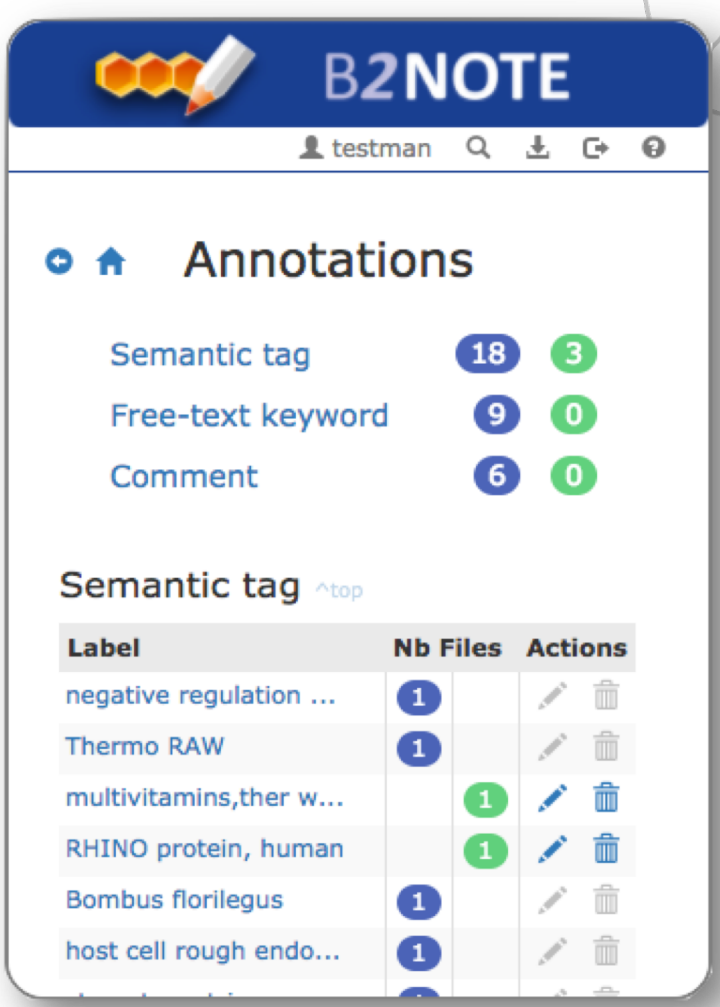




EUDAT Collaborative Data Infrastructure

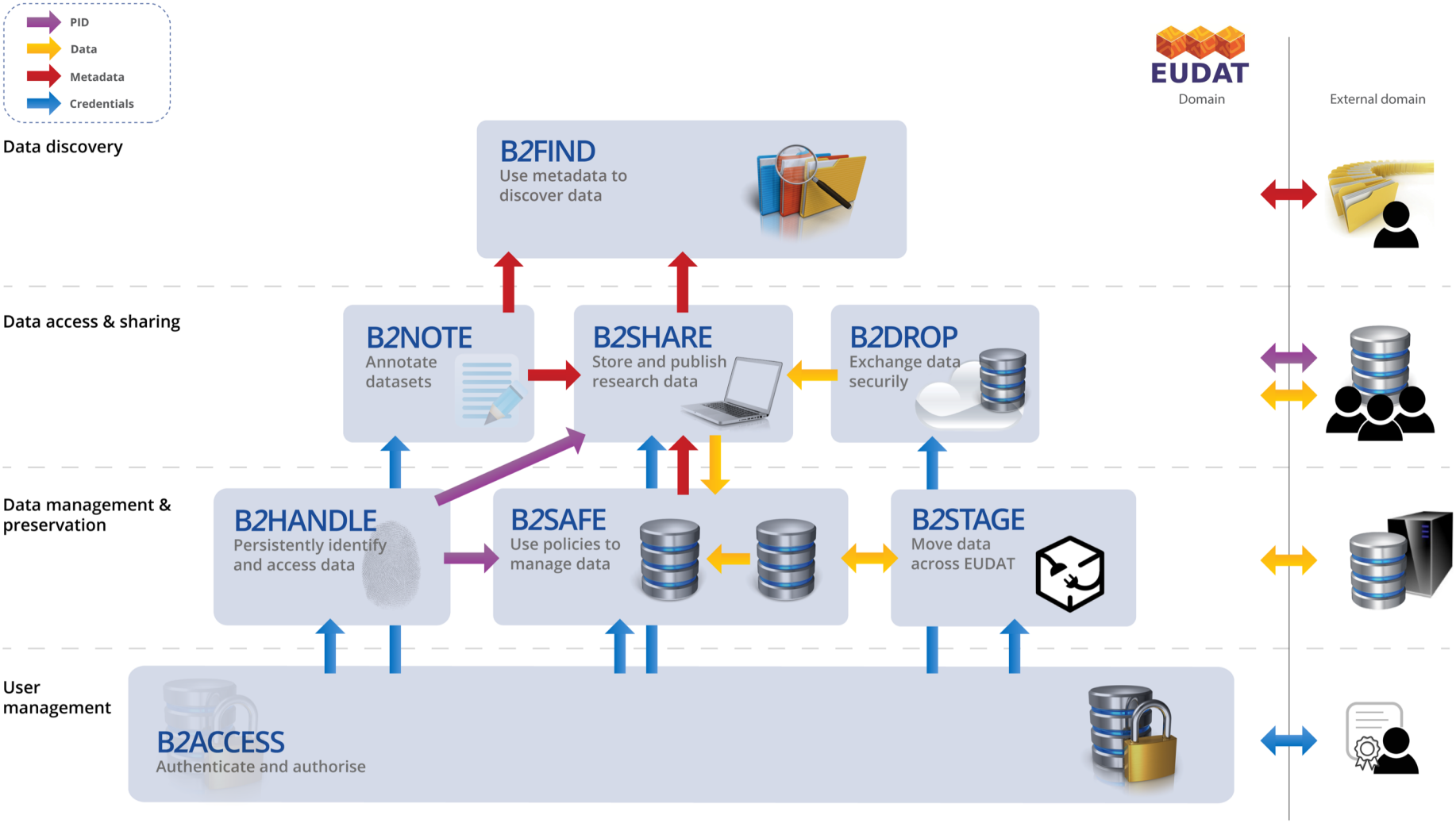
B2NOTE

- **Who**
  - Anyone
- **What**
  - **Enrich** data with Semantic tag, Free-text keyword or comment **without changing the data record**
  - **Share** annotations
  - **Manage** annotations
  - **Integrate with data repositories**
- **Why**
  - **Retrieve and aggregate** heterogeneous files from distributed sources on basis of annotations



https://b2note.eudat.eu/

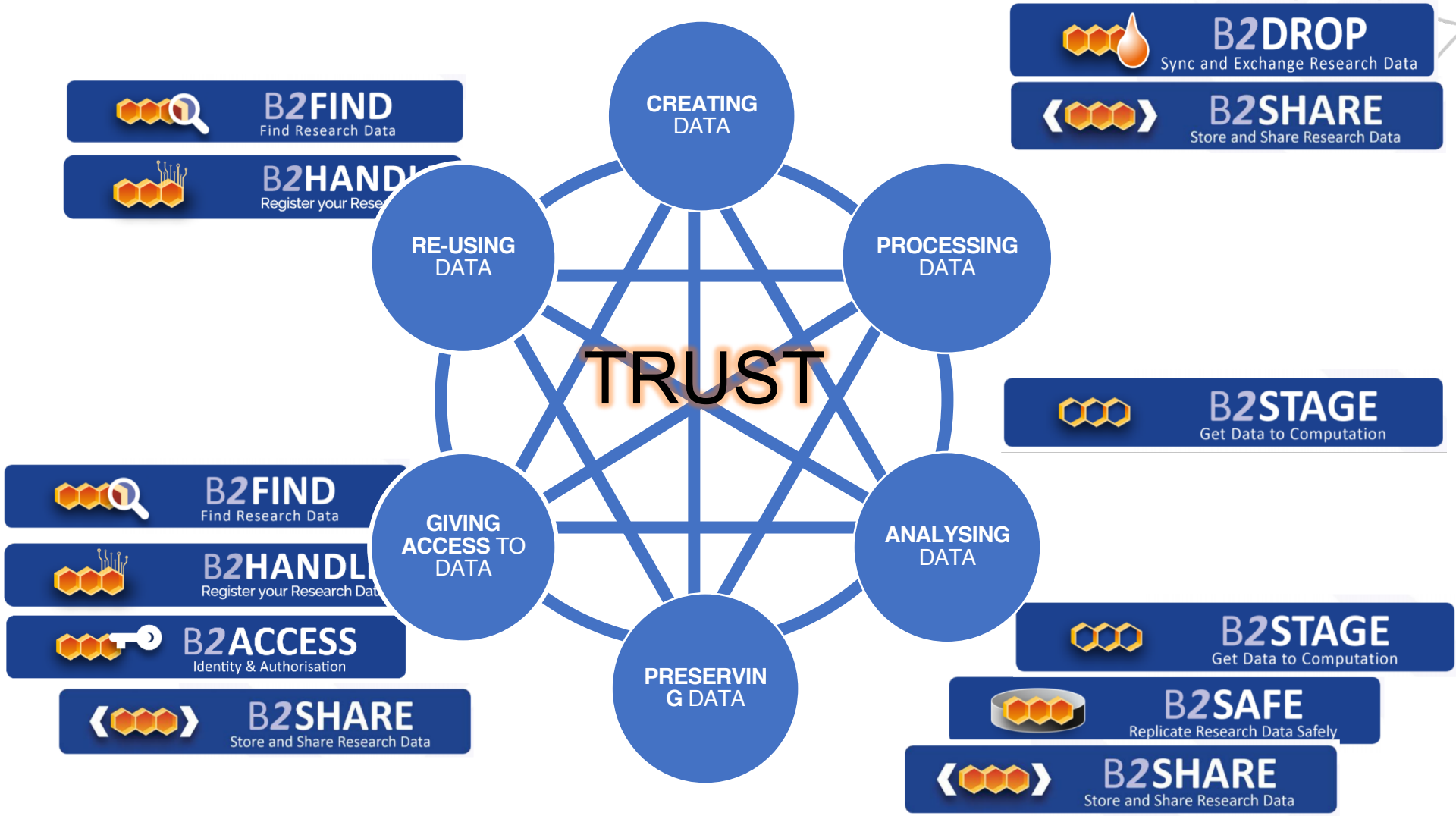
EUDAT B2 Services Diagram



Service Status Overview

Service Component	Development status	Version	Release Level	TRL level	Remark
B2SAFE-CORE	Production	4.1.0	Stable	9	
NEW B2SAFE-DPM	Production	1.2.0	Stable	8	
B2SAFE-METADATA	Proof-of-Concept		Alpha	3	Local metadata store to manage structural metadata. No release defined in GitHub
B2SHARE	Production	2.1.0	Stable	9	
B2DROP	Production	12.0.4	Stable	9	B2DROP version is based on Nextcloud version
NEW B2DROP-B2SHARE bridge	Production	1.0.0	Stable	8	
B2STAGE-GridFTP	Production	1.9.0	Stable	8	
NEW B2STAGE-HTTP	Production	1.0.0	Stable	8	
NEW B2HANDLE	Production	8.1.0	Stable	9	B2HANDLE version is based on Handle version.
B2HANDLE library	Production	1.1.1	Stable	8	
NEW B2ACCESS	Production	1.9.6	Stable	9	B2ACCESS version is based on Unity-IDM version
B2FIND	Production	2.3.2	Stable	9	
NEW B2NOTE	Production	1.0.0	Stable	8	
GEF	Pilot		Beta	6	
DATA DISTRIBUTION	Proof-of-Concept		Alpha	3	
WORKSPACE	Proof-of-Concept	0.4	Alpha	4	Prototype of the HTTP API for workspaces has been released.

EUDAT Services & Data Life Cycle



EUDAT & FAIR

EUDAT Collaborative Data Infrastructure

EUDAT & THE FAIR DATA PRINCIPLES

"Everybody wants to play FAIR, but how do we put the principles into practice?"

EUDAT's vision is data shared and preserved across borders and disciplines and its mission is to enable data stewardship within and between European research communities through the EUDAT Collaborative Data Infrastructure. In 2014 the FAIR guiding principles for individual datasets were formulated by a group of different stakeholders.

These research data principles are widely used now by all possible stakeholders in research data management and are part of the European Commission's data management plans. EUDAT's suite of data management services supports researchers and research communities to ensure their data is FAIR compliant.

Findable

Assign persistent IDs, provide rich metadata, register in a searchable resource...

Interoperable

Use formal, broadly applicable languages, use standard vocabularies, qualified references...

Accessible

Retrievable by their ID using a standard protocol, metadata remain accessible even if data aren't...

Reusable

Rich, accurate metadata, clear licences, provenance, use of community standards...

https://www.force11.org/fairprinciples

EUDAT & FAIR

EUDAT Collaborative Data Infrastructure

At EUDAT we are working to make our services FAIR

EUDAT & FINDABLE

- **B2FIND**: multi-disciplinary metadata catalogue
- **B2HANDLE**: policy-based prefix & PID management
- **B2SHARE**: research data repository
- **B2SAFE**: policy-driven data management

EUDAT & ACCESSIBLE

- **B2STAGE**: data staging service
- **B2SHARE**: research data repository
- **B2SAFE**: policy-driven data management
- **B2NOTE**: research data annotation

EUDAT & INTEROPERABLE

- **B2HANDLE:** policy-based prefix & PID management
- **B2STAGE:** data staging service
- **B2SHARE:** research data repository
- **B2SAFE:** policy-driven data management
- **B2FIND:** multi-disciplinary metadata catalogue

EUDAT & REUSABLE

- **B2SHARE:** research data repository
- **B2SAFE:** policy-driven data management
- **B2NOTE:** research data annotation

User Documentation

EUDAT Collaborative Data Infrastructure

Engage

For Community Decision-Makers & Data Managers

- EUDAT Primer

Services

- B2FIND
- B2STAGE
- B2SAFE
  - What is B2SAFE
  - Using B2SAFE
  - Joining B2SAFE
- B2HANDLE
- B2SHARE
- B2DROP
- Use B2HOST
- Join B2HOST

Tools

- License Selector
- Monitoring information

Deploy

For Systems and Support Engineers

- EUDAT Primer

Services

- B2FIND Integration
- B2STAGE Administration
- B2SAFE Configuration
  - iRODS Deployment
    - MPI-PL, SURFSara, RZG iRODS Zone Federation
    - The dCache to iRODS connection at SURFSara
- B2HANDLE for Communities
- B2SHARE Deployment
- B2ACCESS Management
- B2ACCESS Service Integration

Tools

- Monitoring for Operators
- Resource Coordination Tool
- Site and Service Registry Administration

Use

For Researchers and End-Users

- EUDAT Primer

Services

- B2FIND Usage
- B2STAGE
- B2SHARE
  - B2SHARE Usage
  - B2SHARE API
- B2DROP
  - Publish from B2DROP to B2SHARE
- B2HANDLE for end-users
- B2ACCESS Usage

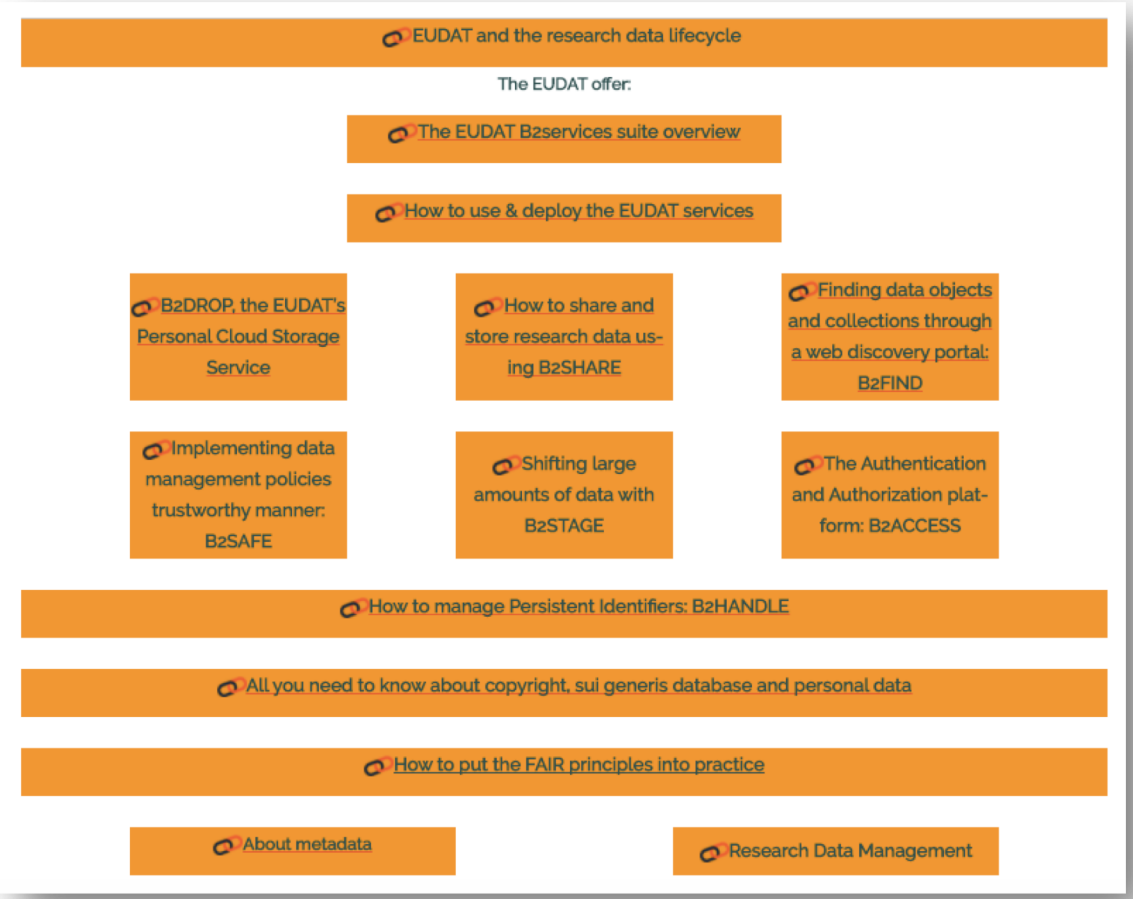
Tools

- License Selector
- Monitoring information

- Total 33 documents maintained and revised
- 3 levels of documentation:
  - Engage: for Community decision-makers and data managers
  - Deploy: for system and support engineers
  - Use: for researchers and end users
- Participation from community experts

https://eudat.eu/services/userdoc

Training Material



- Total of 14 training modules developed and maintained
- Hands-on training environments for:
  - B2SAFE
  - B2SHARE
  - B2FIND
  - B2HANDLE
  - B2NOTE

https://eudat.eu/training - https://github.com/EUDAT-Training

EUDAT Collaborative Data Infrastructure

Use Cases and Hands-on examples

Overview

- Use cases
  - CompBioMed
  - SeaDataCloud
  - CLARIN
- Demos
  - B2DROP -> B2SHARE publication workflow
  - B2FIND -> B2SHARE discovery and download
  - B2DROP -> CLARIN Switchboard example

Safe data replication with B2SAFE

- CompBioMed is a European commission H2020 funded Centre of Excellence
- Focus on the use and development of computational methods for biomedical applications.
- Data-intensive research
- More than 40 international and associate partners

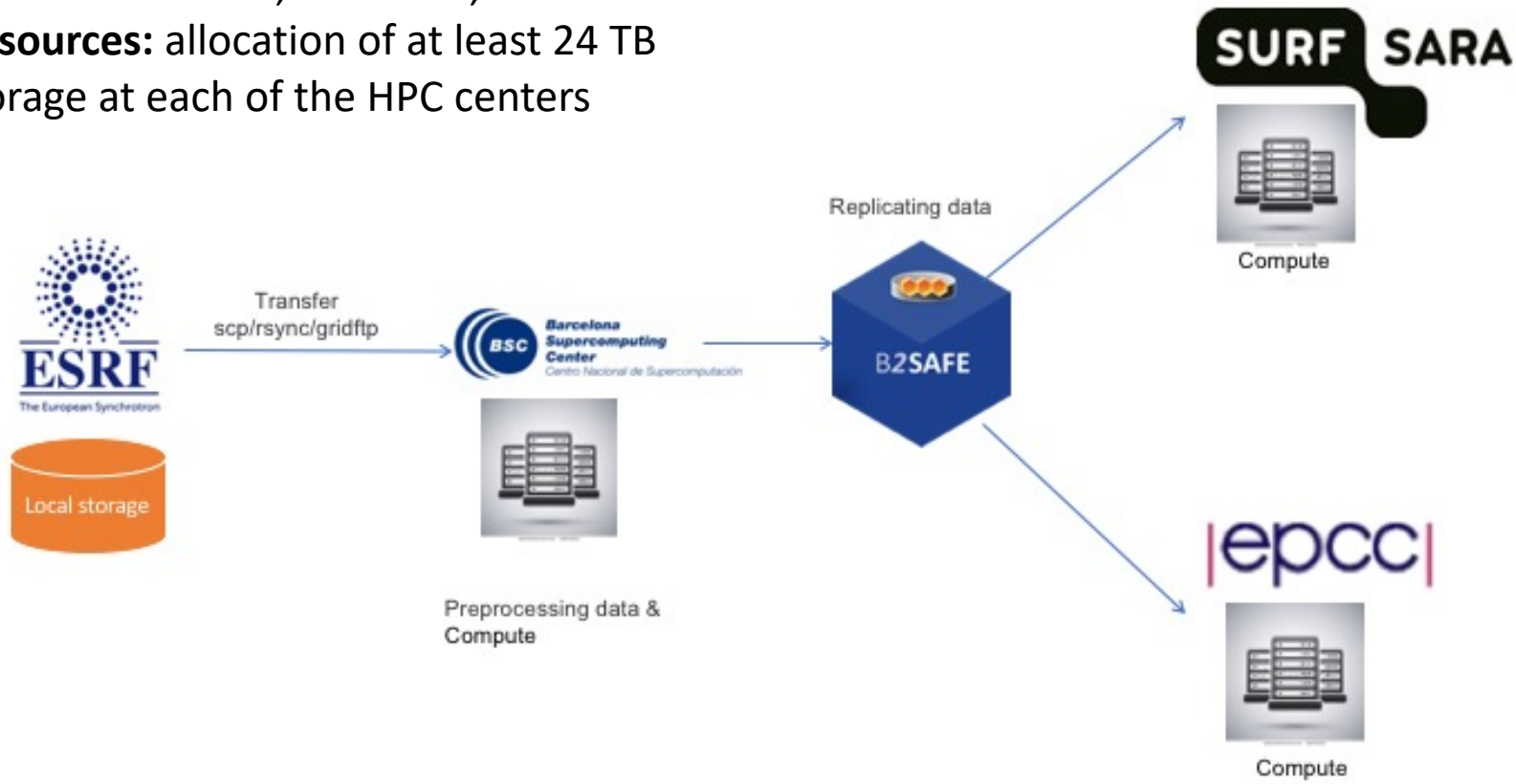
Safe data replication and large data transfer is one of the major requirements within the CompBioMed community

https://www.compbiomed.eu/

CompBioMed: Data Pipeline

Resources

Service: EUDAT B2SAFE service
HPC Centers: BSC, SURFsara, EPCC
Resources: allocation of at least 24 TB storage at each of the HPC centers



EUDAT Collaborative Data Infrastructure

SeaDataCloud

- SeaDataNet consortium operates a state-of-the-art pan-European infrastructure to manage high quality ocean and marine data
- SeaDataCloud is the third proposal, after SeaDataNet and SeaDataNet2

  (https://www.seadatanet.org/About-us/SeaDataCloud)
- Duration: 2016 - 2019
- Aim:
  - To advance SeaDataNet service and increase their usage by adopting cloud and HPC technology
- EUDAT CDI:
  - Leverage EUDAT CDI infrastructure for long-term digital preservation and curation provide unified data access
  - 5 partners: DKRZ, CINECA, CSC, GRNET, STFC
  - B2 services: B2DROP, B2SHARE, B2SAFE, B2HOST, B2STAGE, B2FIND and B2ACCESS

SeaDataCloud: the challenge

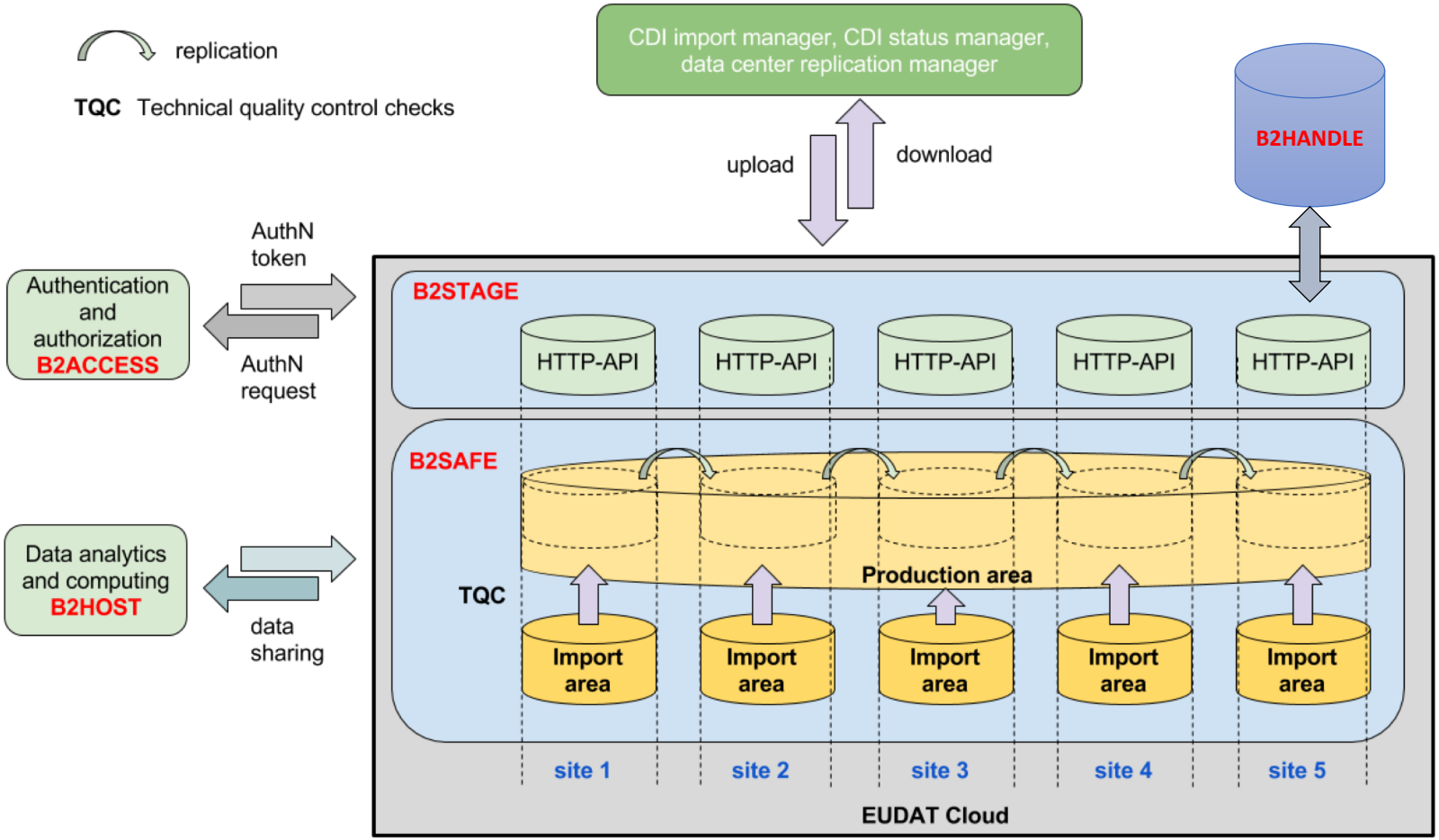
- The SeaDataNet portal (CDI: Common Data Index) collects only part of the data produced by more than a hundred marine research institutions.

- The others are stored locally from the institutions and offered to the users after a request via email. They are made accessible via a temporary web service endpoint.

- The quality checks are performed by the local institutions, without any central mechanism, therefore the risk of inconsistencies and duplications is high.

- There is not a Virtual Research Environment, but a set of desktop and web applications , independent from each other. The user is forced to upload the data set that she wants to analyze and to download the result: there is not a shared data space, neither there is a personal one.

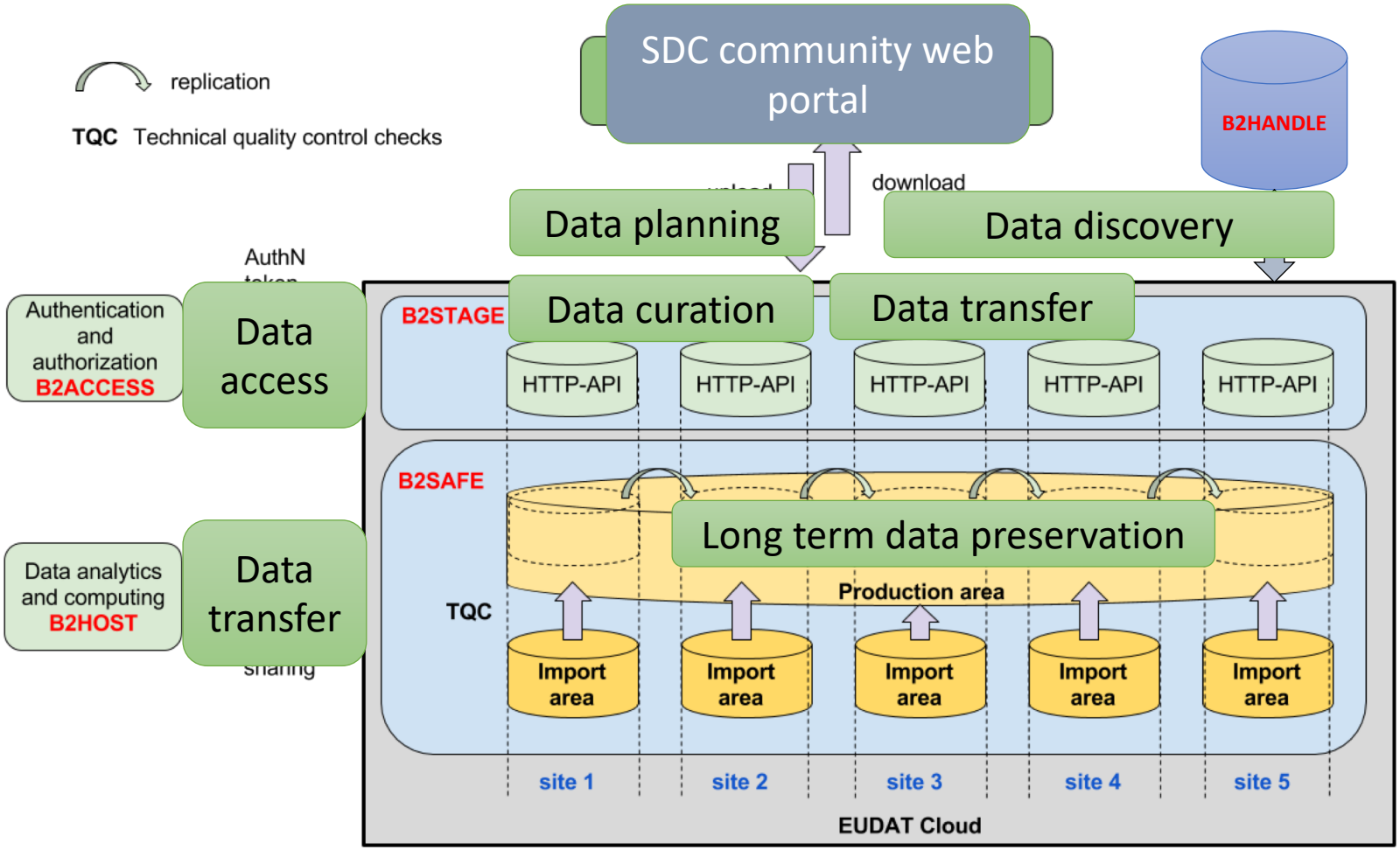
SeaDataCloud: B2SAFE and B2STAGE



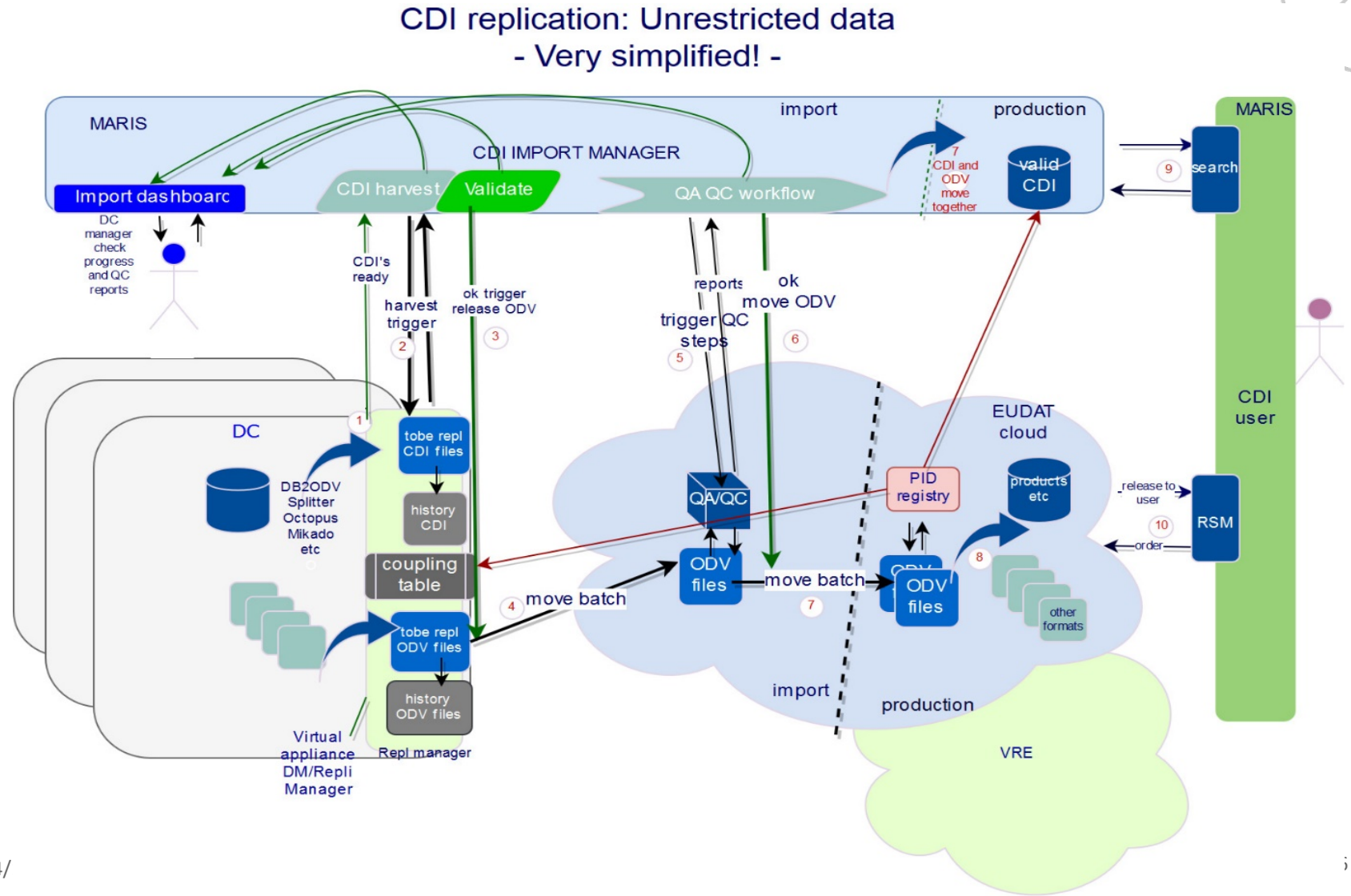
SeaDataCloud: The Solution

- B2SAFE and B2STAGE services are hidden behind the community web portal (CDI) which takes care to manage user and community specific metadata registration (**DATA DISCOVERY**).
- Each of the five EUDAT data centers offers a B2SAFE instance federated with the others.
- Each data center provides two storage areas:
  - one for the ingestion of the new data uploaded by the data producers, which are the hundreds of marine science institutions of SeaDataNet (**DATA TRANSFER**);
  - one for the production ready data, which have been validated by the data manager through the community web portal.
- The community web portal triggers quality check workflows on the B2SAFE and B2HOST side (**DATA PLANNING, DATA CURATION**).
- Once moved into the production area, the data are replicated following a star pattern: each replica has the same master copy. And a B2HANDLE PID is associated to them (**LONG TERM DATA PRESERVATION**)
- Data can then be shared with applications running on the B2HOST environment (**DATA TRANSFER**)

SeaDataCloud: B2SAFE solution



The Real SeaDataCloud Data Flow

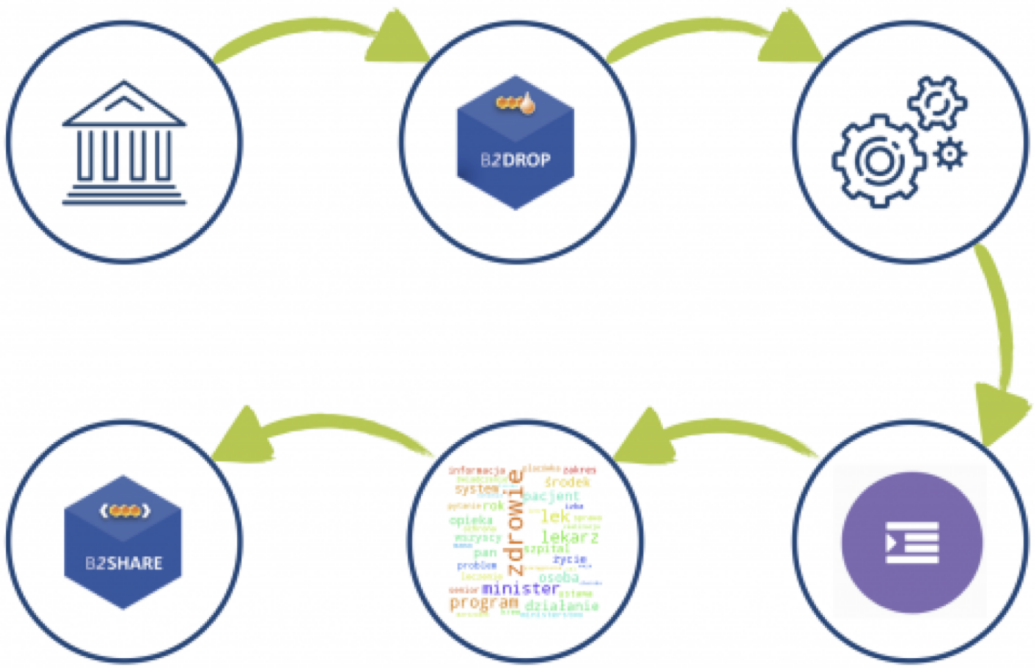


CLARIN Use Case

- CLARIN is the European Research Infrastructure for Language Resources and Technology. It makes digital language resources and language technology available to researchers from all disciplines but especially the Humanities and Social Sciences
- CLARIN is a partner in EOSC-hub and
  - Potentially contributes a huge amount of metadata for language resources and services to B2FIND.
  - Provides thematic service(s) : CLARIN Metadata Infrastructure
    - Virtual Language Observatory (VLO)
    - Virtual Collection Registry (VCR)
    - Language Resource Switchboard
- These services should be integrated with other EOSC-hub services e.g.
  - CLARIN metadata visible in B2FIND
  - Language Resource Switchboard integrated in B2FIND and B2DROP
  - …

CLARIN Switchboard and B2DROP

- First step in CLARIN/EOSC-hub services integration
- Extensive demo available at https://www.clarin.eu/eosc



Demos

- B2DROP -> B2SHARE publication workflow
- B2FIND -> B2SHARE discovery and download
- B2DROP -> CLARIN Switchboard example

EUDAT Collaborative Data Infrastructure

Thanks to:
Claudio Cacciari
Daan Broeder
Dieter Van Uytvanck
Mark van de Sanden
and other EUDAT colleagues

Please provide your feedback about the training session:

https://www.surveymonkey.com/r/EOSC-hub_week_01