Using EUDAT services to replicate, store, share, and find cultural heritage data

…in PSNC… and beyond…

Maciej Brzeźniak, Norbert Meyer, PSNC
Damien Lecarpentier, CSC

Plan and purpose of presentation

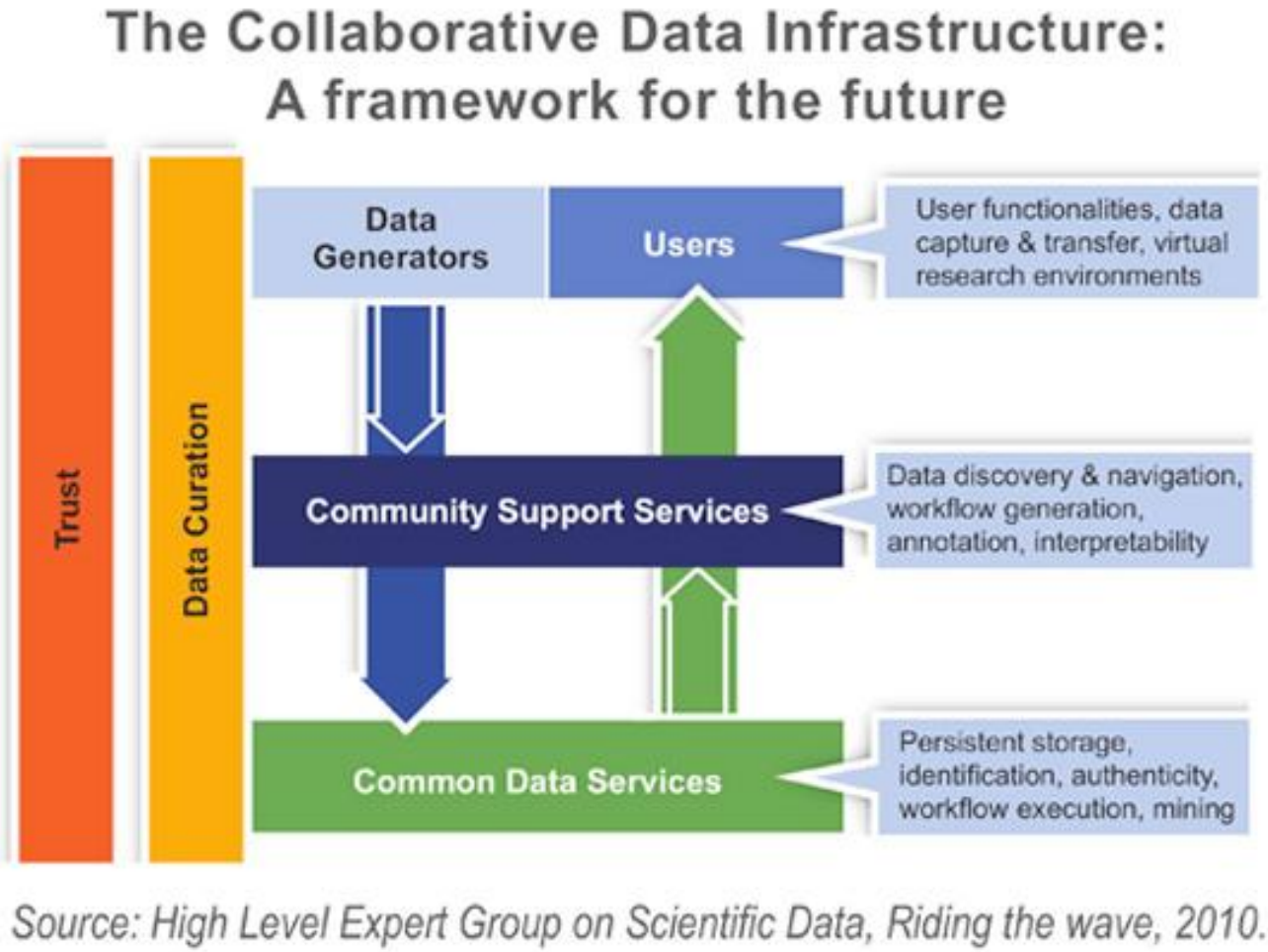
- **Summarize the work done and planned in EUDAT an DCH-RP project**
  - How DCH sector can use e-Infrastructures?
- **Undersdand how selected EUDAT services can be used in DCH domain?**
  - How EUDAT CDI architecture can be integrated with domain-specific services?
  - Overview of current services (Simple Store, Safe Replication)
  - Presentation of integration work in DCH-RP
  - *Discussion on possible future extensions / services*

Some aspects of EUDAT

EUDAT – European Data Infrastructure

- **Vision**: to support a Collaborative Data Infrastructure
- **Aims**:
  - Provide a **sustainable platform** of **technologies**, **tools, services driven by user needs**
  - **Engage users in defining/shaping a platform for shared services**
  - Support **data-intensive, multi-disciplinary research**:
    - Humanities and Social Science: CLARIN
    - But also: earth science (ENES, Earth system modelling; EPOS: European Plane Observing System), ecology (LifeWatch), Virtual Physiological Human (VPH)
  - Deliver **common low-level services** that are required to provide the level of interoperation and trust of data
  - Ensure that the data **infrastructure is robust/scalable** (able to address ‚data tsumami')
  - **Build community/domain-specific services** on top of the common services with participation of users

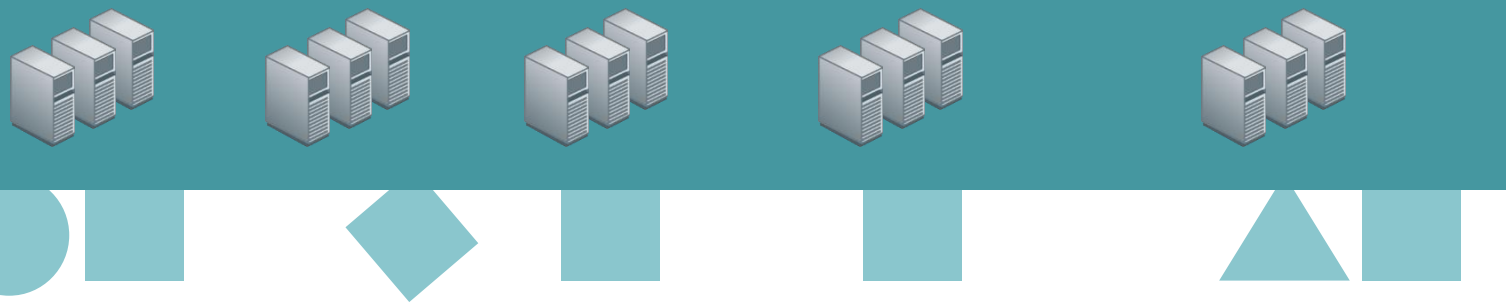
CDI layers
Commons vs community-specific services



Source: High Level Expert Group on Scientific Data, Riding the wave, 2010.

CDI layers vs tools and services in DCH
Integration challenges…

- *High-level services:*
  - **EUDAT simple store**
  - **eCulture Science Gateway**
  - **dLibra dArceo dLab**
  - **Invenio**
  - **Another community solution**



- *Low-level services:*
  - **Local LTS:** long-term storage
  - **EUDAT storage**
    - Safe Replication
  - **Cloud storage**
    - S3
    - CDMI
  - **Grid storage**
    - FTS
    - GridFTP

CDI layers vs tools and services in DCH
Integration challenges…

- *High-level services:*
  - **EUDAT simple store**
  - **eCulture Science Gateway**
  - **dLibra dArceo dLab**
  - **Invenio**
  - **Another community solution**

DCH-RP <-> EUDAT proof of concept: Do they go together?

- *Low-level services:*
  - **Local LTS:** long-term storage
  - **EUDAT storage**
    - Safe Replication
  - **Cloud storage**
    - S3
    - CDMI
  - **Grid storage**
    - FTS
    - GridFTP

EUDAT: services for DCH domain

- **Services covered in the presentationis**
  - **Simple Store service:**
    - enable researchers and scientists to upload, store and share date
    - designated for the „long tail of data”
  - **Safe Replication:**
    - Allow communities to replicate data to selected EUDAT data centers
    - Automated replication (iRODS), PID registration (EPIC)
  - **Data Staging:**
    - Staging data from user community premises/systems (iRODS)
    - to computing systems, e.g. PRACE’s HPC centres (GridFTP, FTS)
  - **Metadata service:**
    - Joint metadata domain for all EUDAT data centres
    - Searchable catalogue covering all data stored within EUDAT
  - **AAI:**
    - Provide a solution for a working AAI system in a federated

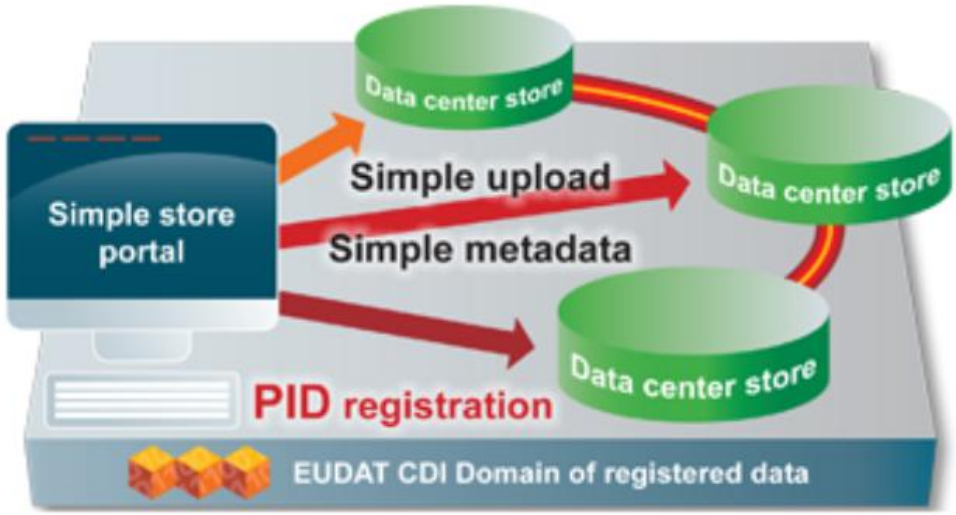
EUDAT for DCH
Simple Storage Service

Simple Storage Service:

- Address the issues of small user groups and indivitual users
- Provides solution for "long tail data": often stored on laptops and departmental servers

Functionality:

- alowing registered users to **upload typical** data objects into the EUDAT store
- enabling users to **share** such **objects** and collections with other researchers.
- lets utilising other EUDAT services
  - Safe Replication
  - PIDs
  - etc.
- May be integrated with AAI



More:
http://www.eudat.eu/simple-store

EUDAT for DCH
Simple Storage Service

Simple Storage Service internals:

- Referred also as Researcher Data Store
- Based on Invenio:
  - Developed by CERN
  - http://invenio-software.org/
- Storage backend:
  - Disk
  - iRODS (EUDAT safe replication)
- Front-end:
  - Developed a new submission portal to invenio

EUDAT for DCH
Simple Storage Service

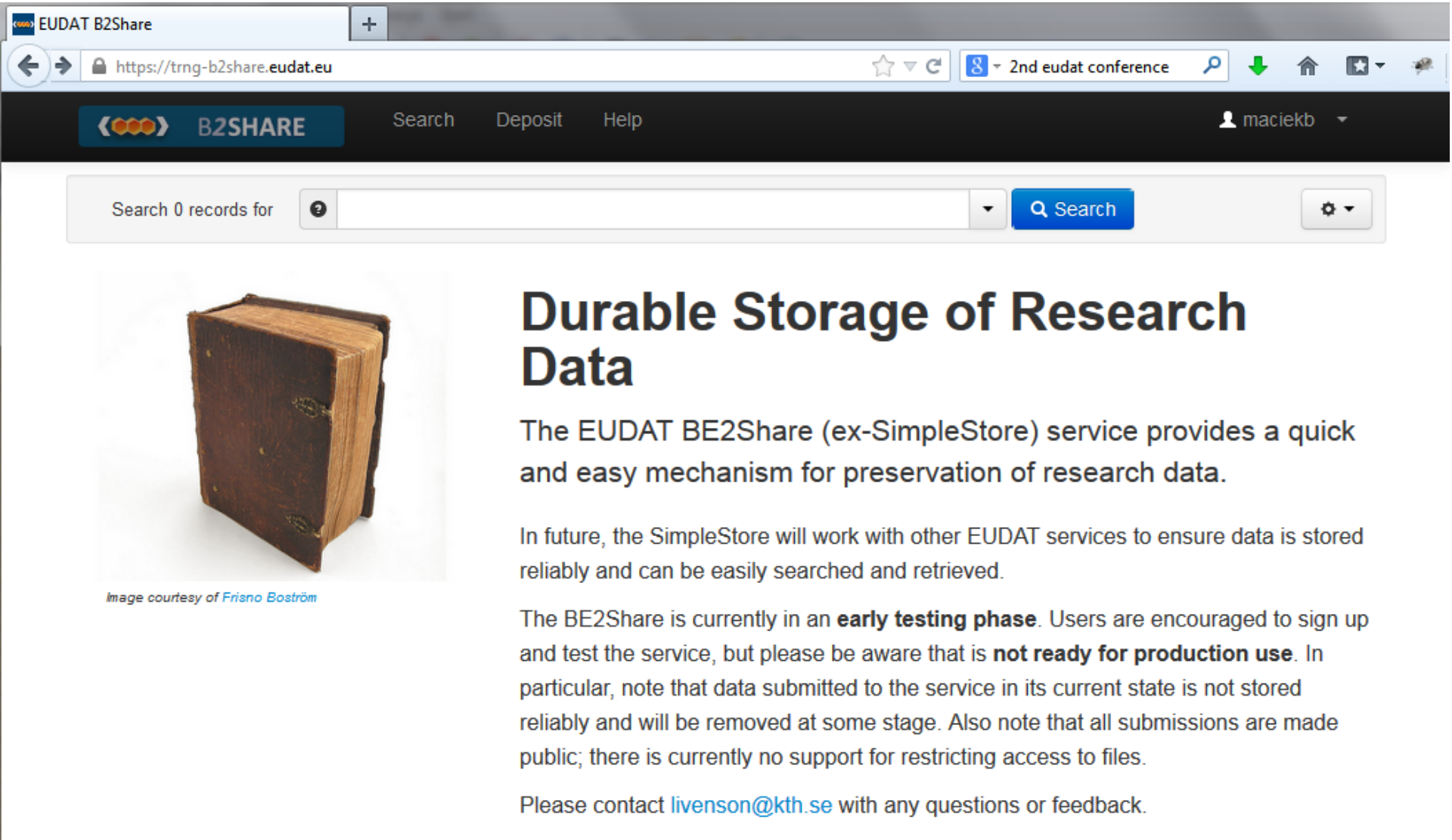


Image courtesy of Frisno Boström

Durable Storage of Research Data

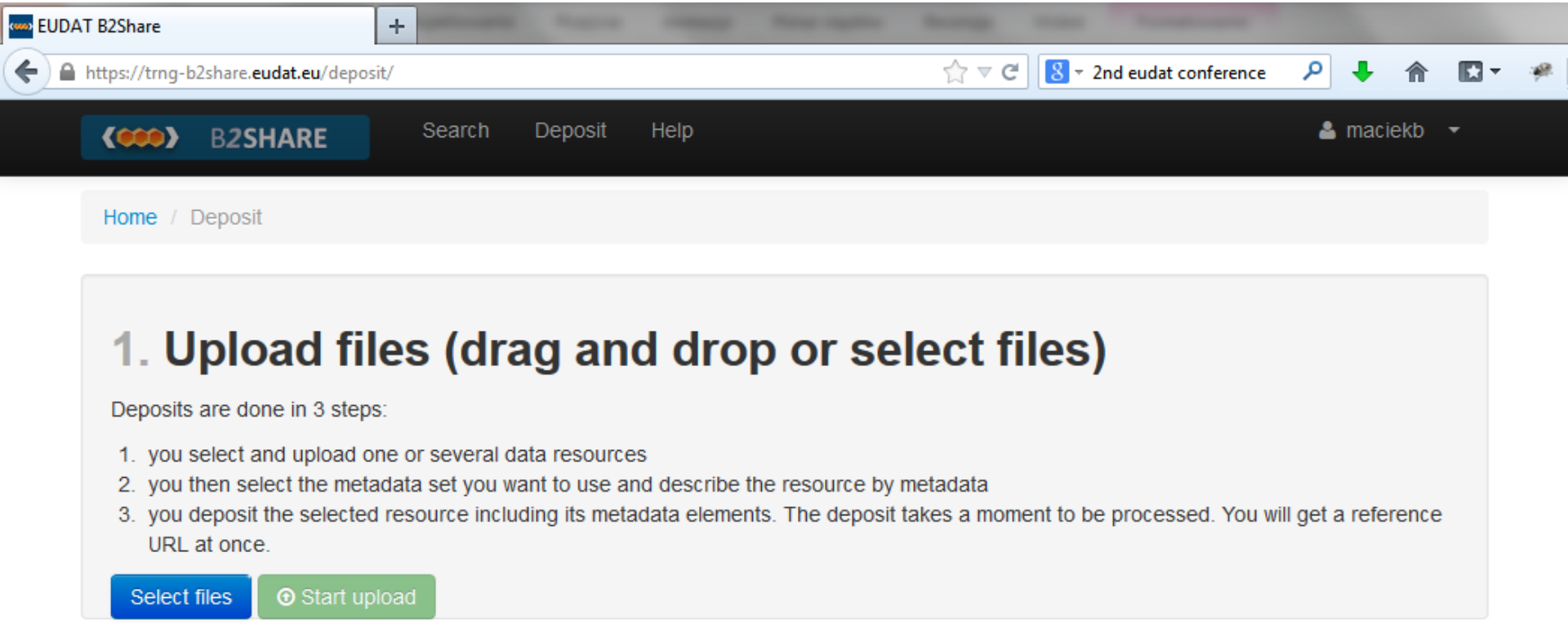
The EUDAT BE2Share (ex-SimpleStore) service provides a quick and easy mechanism for preservation of research data.

In future, the SimpleStore will work with other EUDAT services to ensure data is stored reliably and can be easily searched and retrieved.

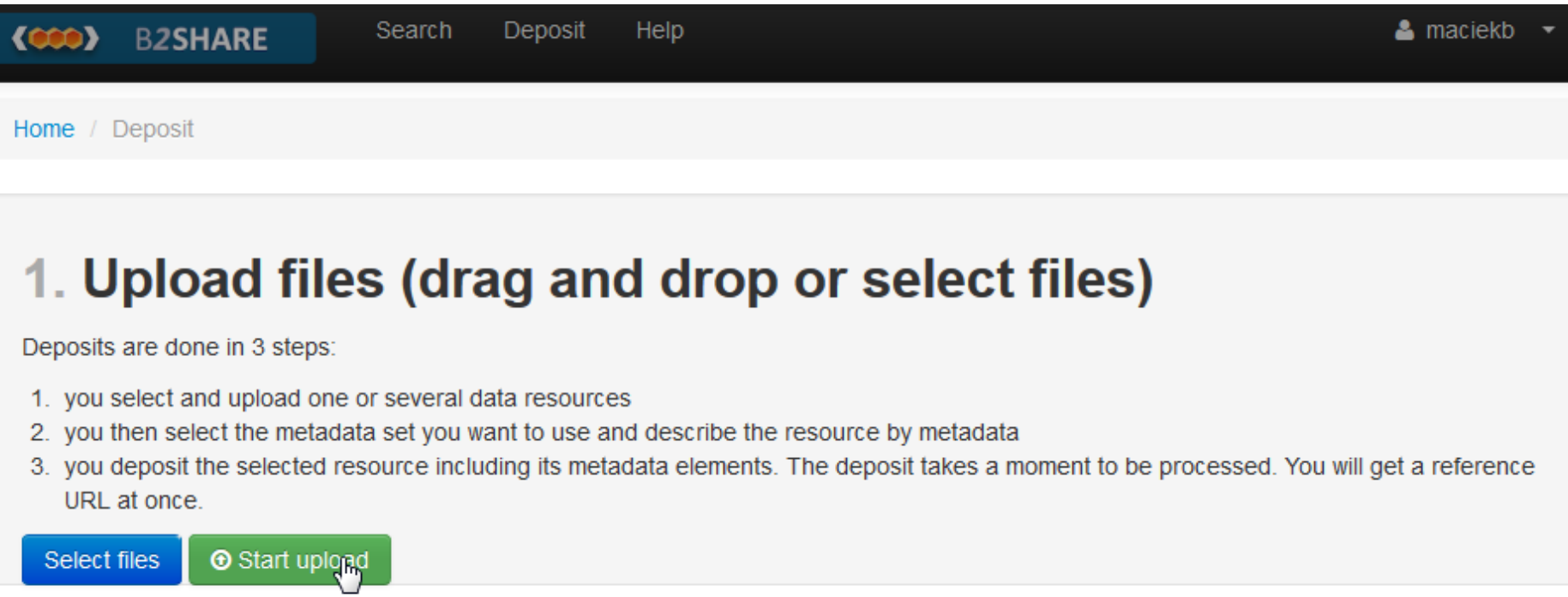
The BE2Share is currently in an **early testing phase**. Users are encouraged to sign up and test the service, but please be aware that is **not ready for production use**. In particular, note that data submitted to the service in its current state is not stored reliably and will be removed at some stage. Also note that all submissions are made public; there is currently no support for restricting access to files.

Please contact livenson@kth.se with any questions or feedback.

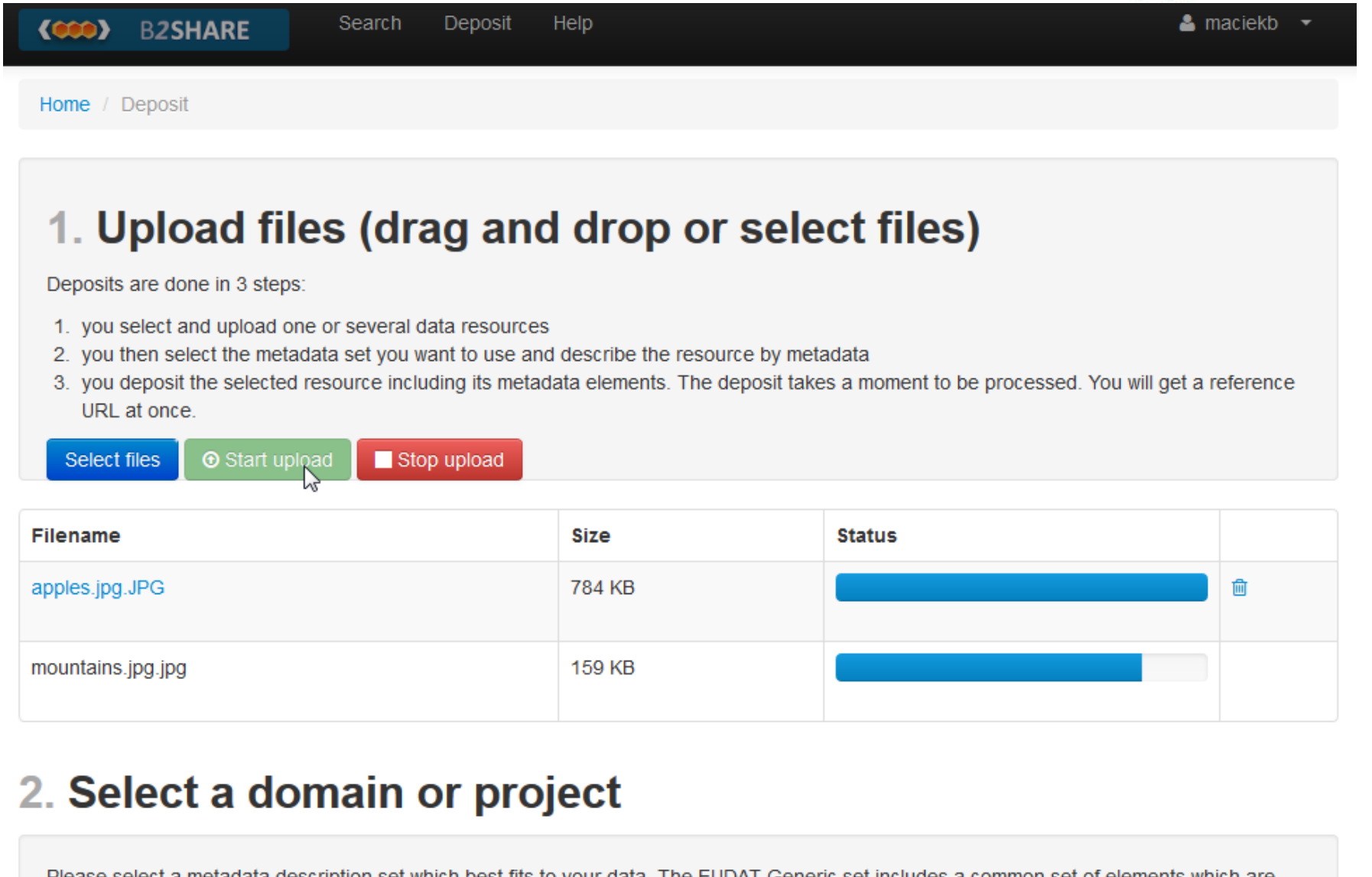
EUDAT for DCH
Simple Storage Service



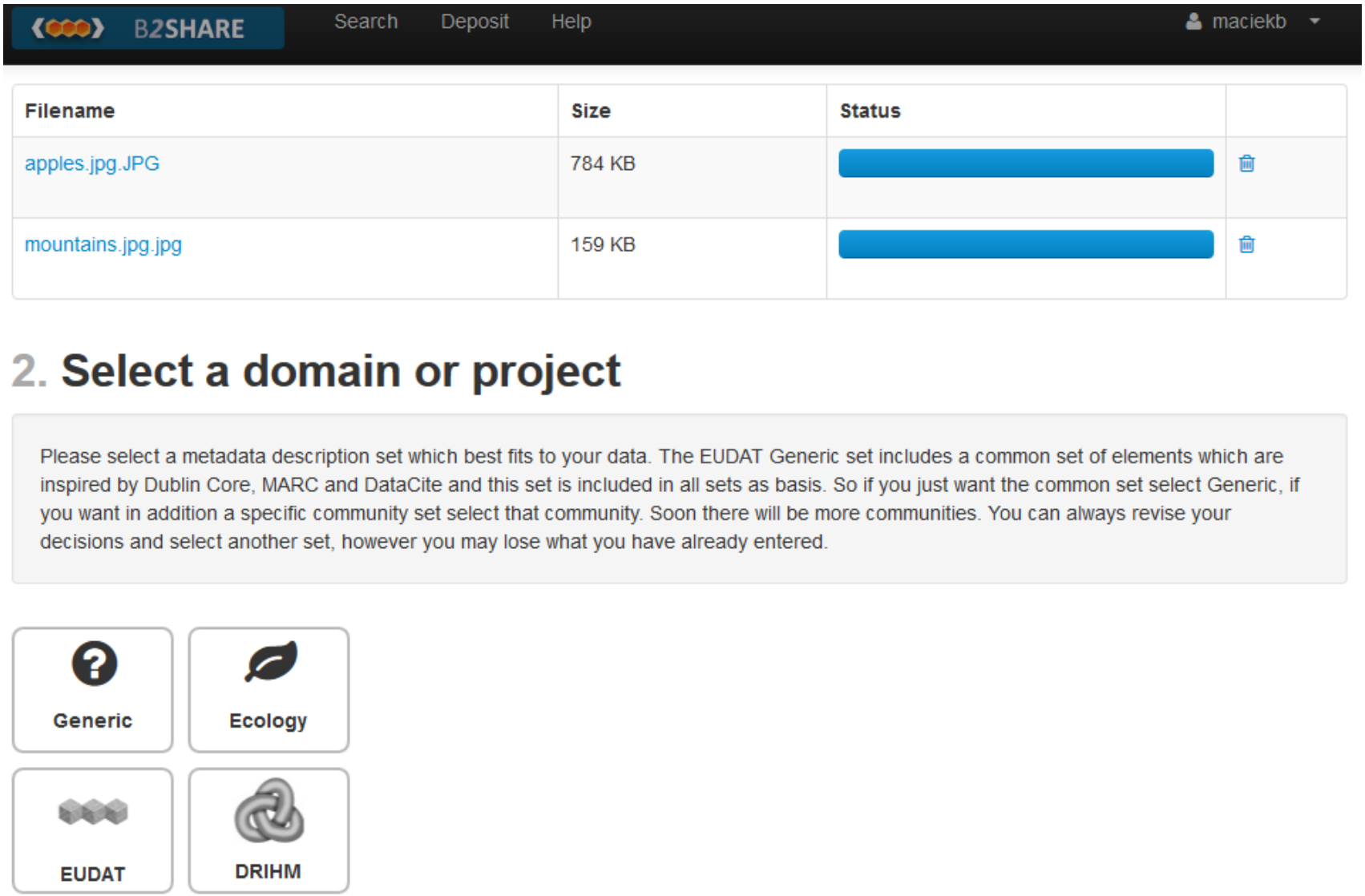
EUDAT for DCH
Simple Storage Service



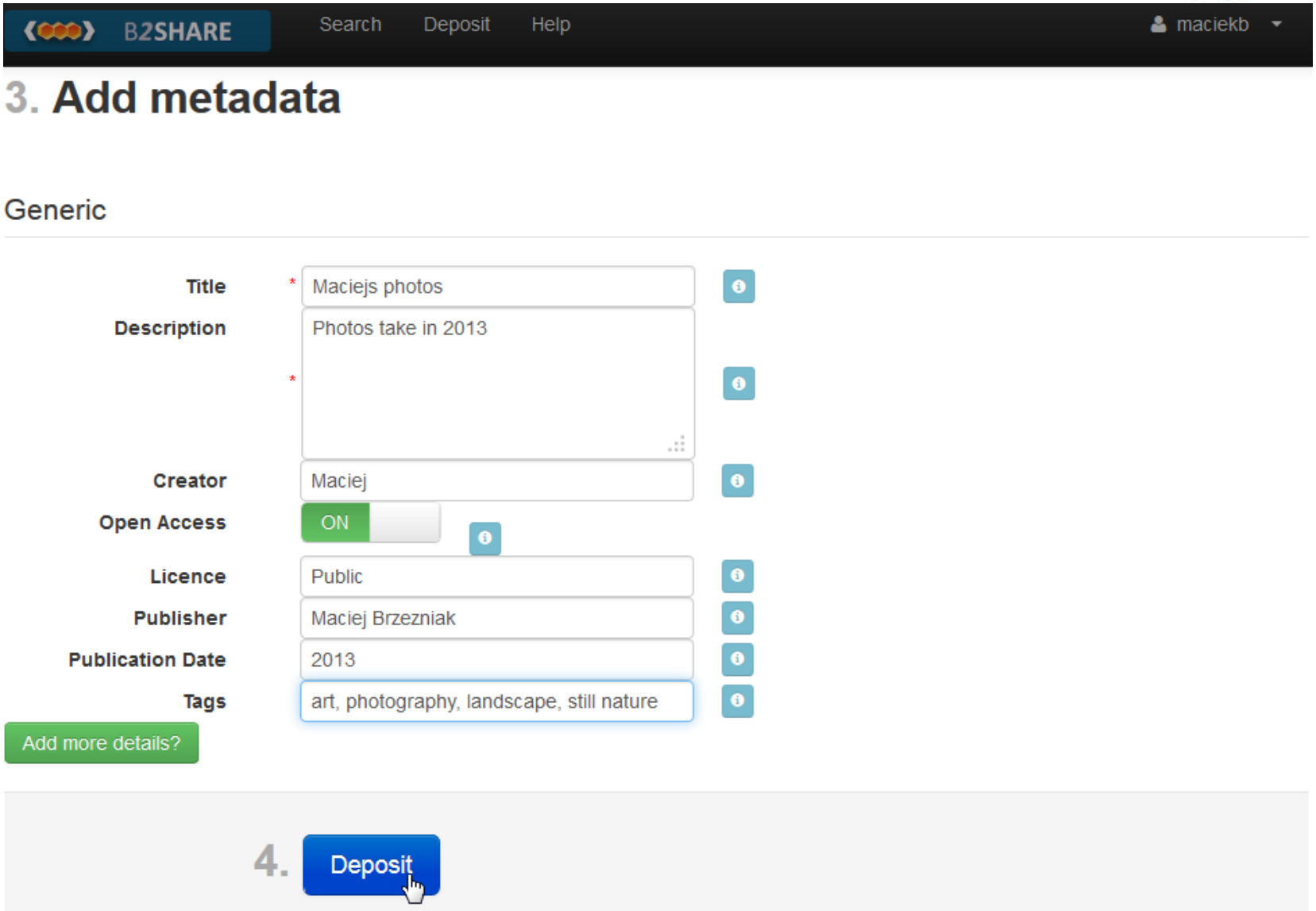
EUDAT for DCH
Simple Storage Service



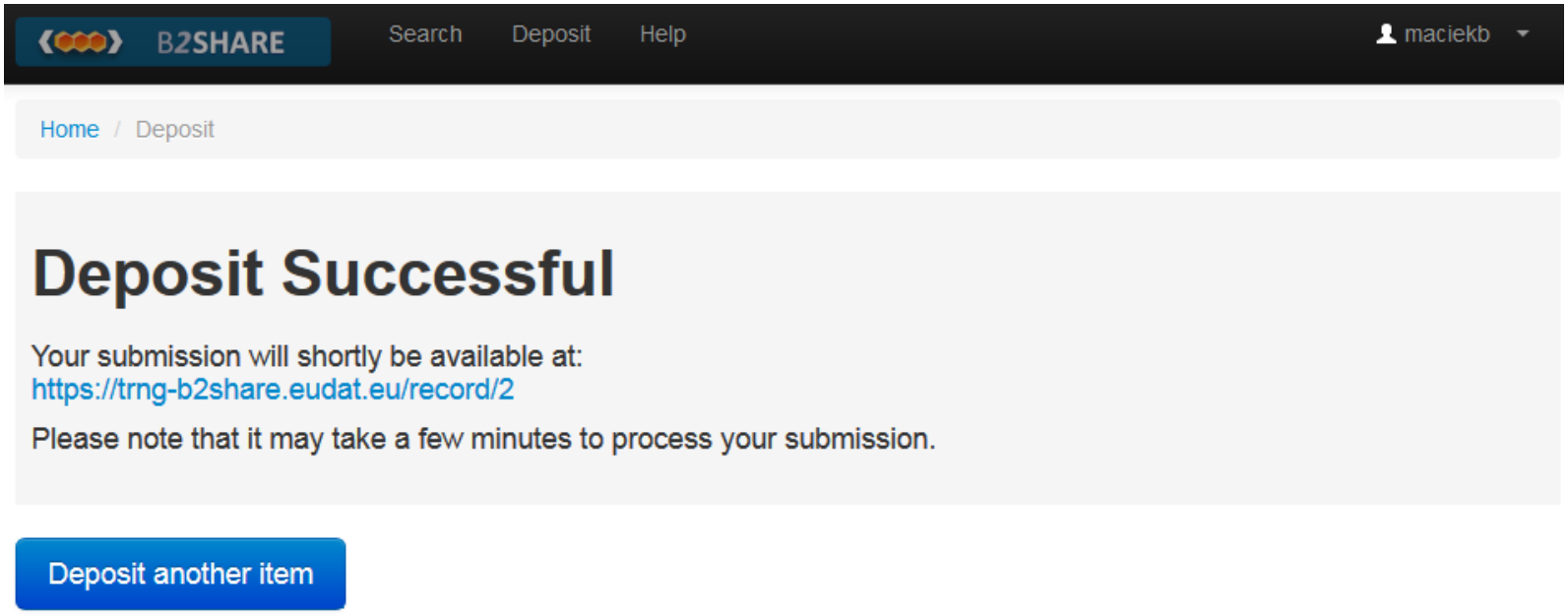
EUDAT for DCH
Simple Storage Service



EUDAT for DCH
Simple Storage Service



EUDAT for DCH
Simple Storage Service



B2Share Search Deposit

Search 61 records for

Search

Add to search AND OR AND NOT

Author

Title

Report number

Affiliation

Collaboration

Keywords

Journal Any Vol Pg

simple advanced Browse Search

Image courtesy of Frisno Bostn

of Researc

eStore) service provi

vation of research da

her EUDAT services to ens

trieved.

g phase. Users are encou

hat is **not ready for produ**

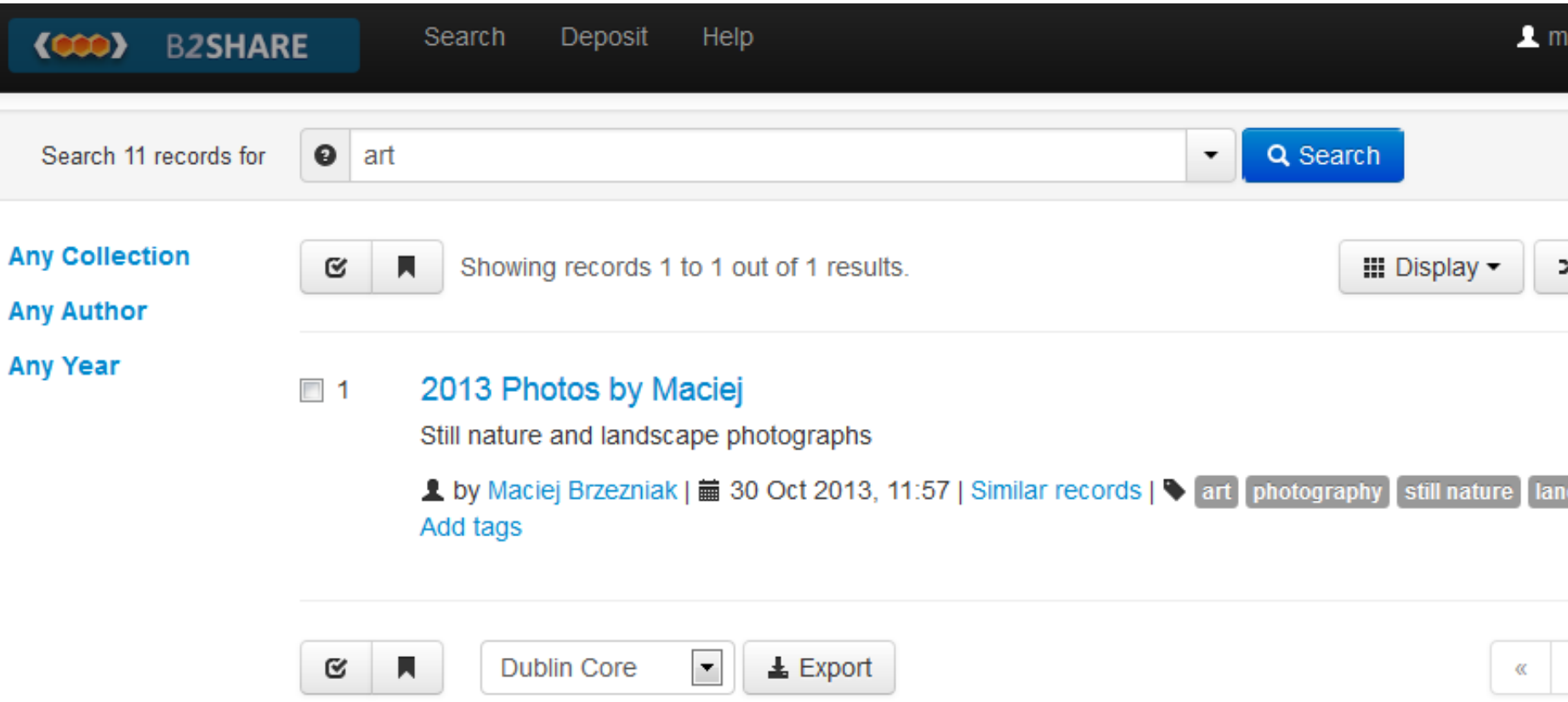
ervice in its current state is

. Also note that all submiss

tricting access to files.

uestions or feedback.

EUDAT for DCH
Simple Storage Service



2013 Photos by Maciej

Maciej Brzezniak

Maciej Brzezniak

Abstract: Still nature and landscape photographs

Keyword(s): art ; photography ; still nature ; landscape

The record appears in these collections:
eudat

Add tags

Back to search

Files

Name	Date	Size	
mountains.jpg.jpg	30 Oct 2013	163.3 kB	Download
apples.jpg.JPG	30 Oct 2013	803.2 kB	Download

Export

BibTeX, MARC, MARCXML, DC, EndNote, NLM, RefWorks

Metadata

Checksum:	3f19aeac3c40de94656b9f3db978b338389870466a9bf223f6aa2e2dc4bee458
PID:	ad99da14-4149-11e3-bc23-14feb57d12b9
Publication:	Maciej Brzezniak
Licence:	Public
Uploaded by:	maciekb@man.poznan.pl
Domain:	eudat

EUDAT for DCH
Safe Replication Service

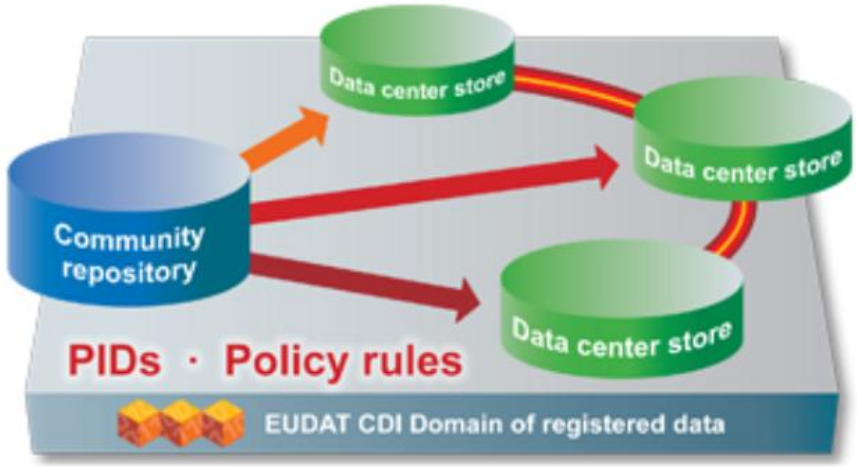
DCH RP — Digital Cultural Heritage Roadmap for Preservation

Safe Replication Service:

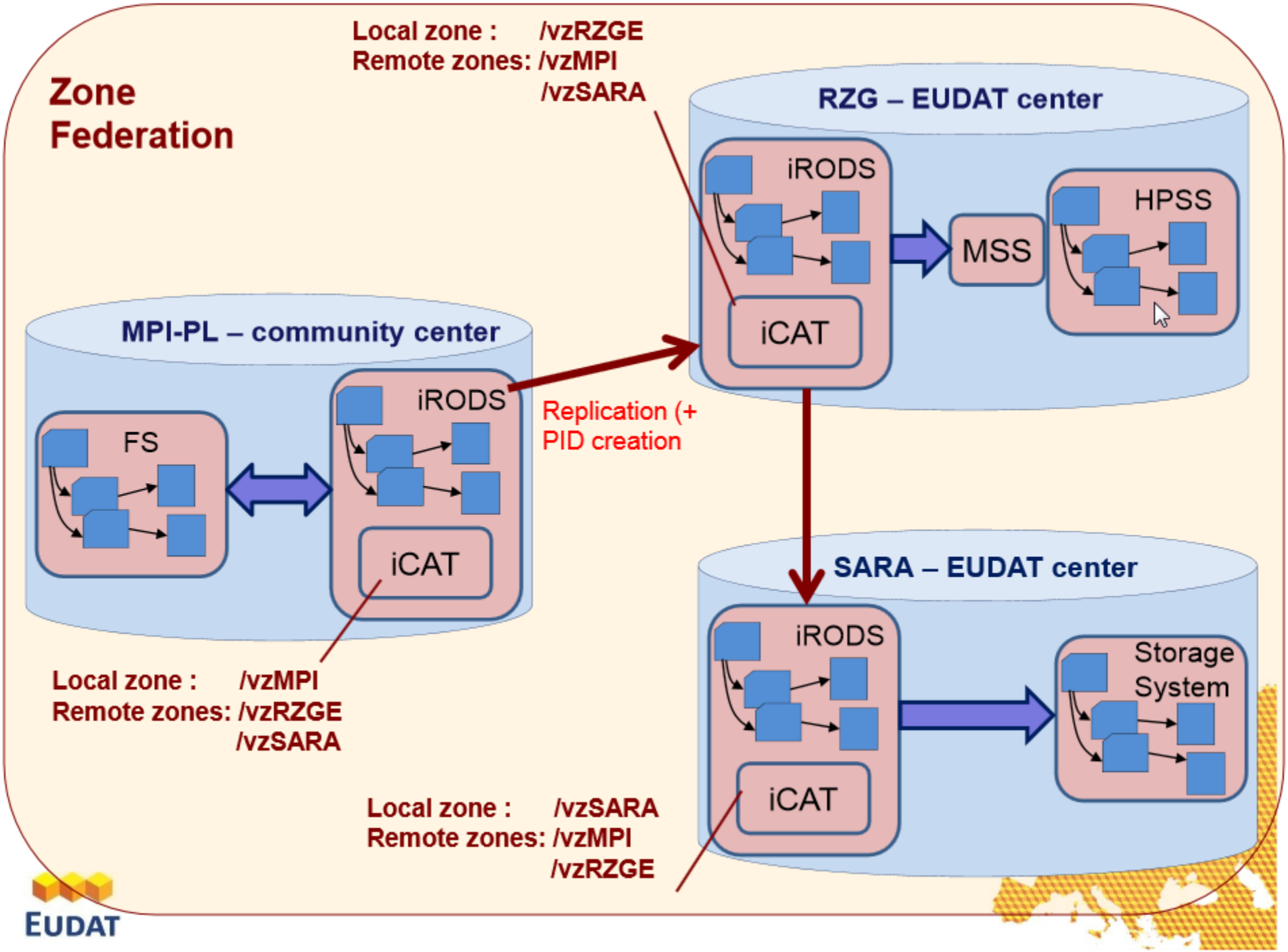
- Allows communities to **replicate data to EUDAT data centers**
- Can be **integrated with portals and community tools**

Functionality:

- Ingested file/data are:
  - automatically **replicated** to many data centres
  - **get persistent identifiers registered** (PIDs based on EPIC)
- Various interfaces supported
  - iRODS: icommand, API,
  - WebDAV, GridFTP
- Can replicate data on top of various different data stores:
  - Disks, tapes, HSMs
  - Clouds (e.g. S3)



More:
http://www.eudat.eu/safe-replication

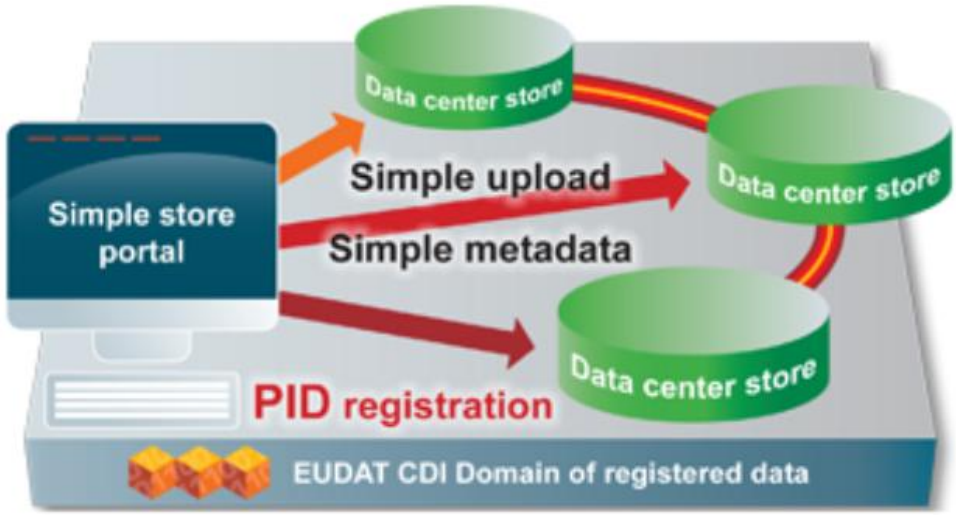

Zone Federation
Local zone : /vzRZGE
Remote zones: /vzMPI
/vzSARA
RZG – EUDAT center
iRODS
MSS
HPSS
iCAT
MPI-PL – community center
FS
iRODS
iCAT
Replication (+ PID creation
Local zone : /vzMPI
Remote zones: /vzRZGE
/vzSARA
SARA – EUDAT center
iRODS
Storage System
iCAT
Local zone : /vzSARA
Remote zones: /vzMPI
/vzRZGE
EUDAT

Integration option 1: EUDAT Simple Store + EUDAT Safe Replication

Support for sharing

Easy deposit & access to the data

Transparent replication of data, persistent storage

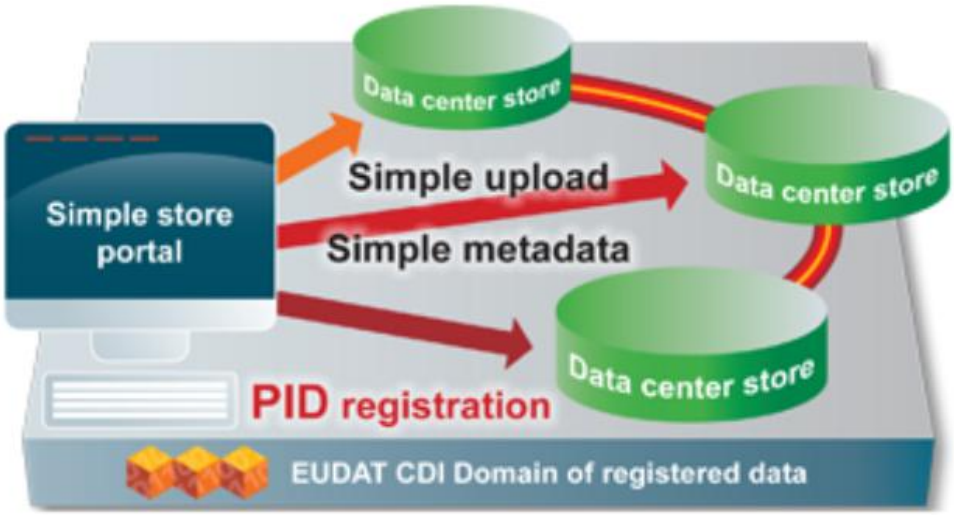


Integration option2: Community portal/solution + EUDAT Safe Replication

Support for sharing

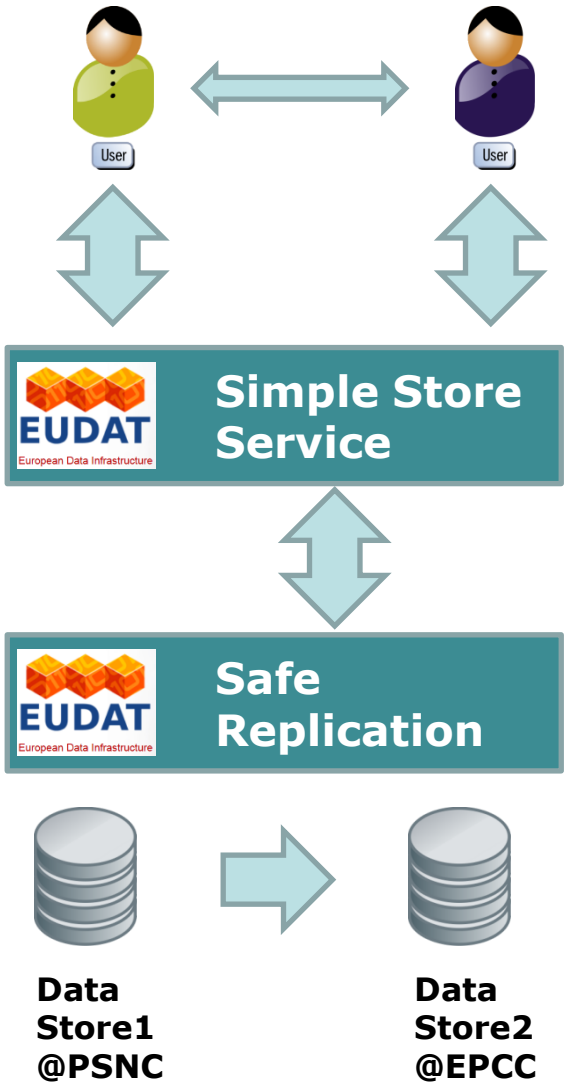
Storage & access typical for community

Transparent replication of data



Option 1 in practice: EUDAT Simple store + EUDAT Safe Replication

DCH-RP <-> EUDAT PoC 1



Support for sharing

Easy storage & access to the data

Transparent replication of data

Memory institutions:

EUDAT Simple Store

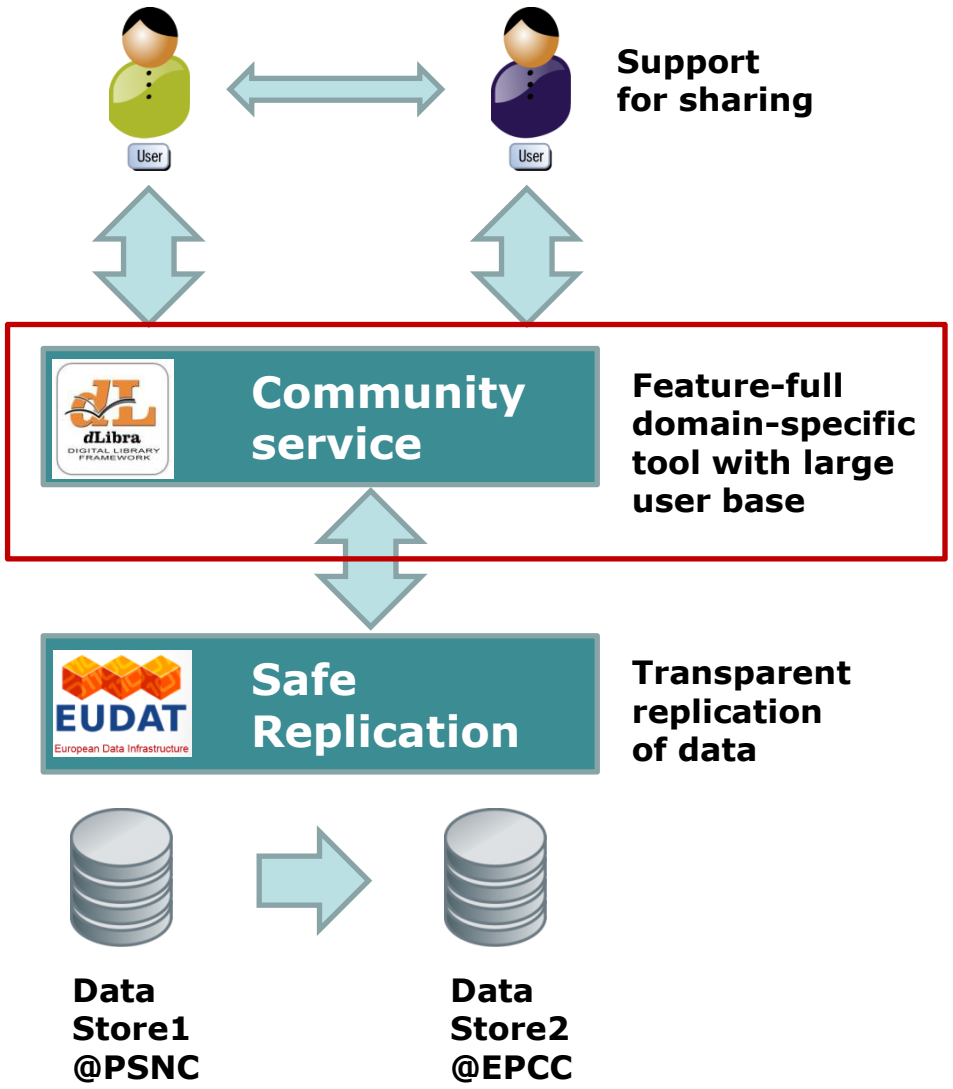
Provided by PSNC

EUDAT Safe replication

Provided by PSNC (Poznan) & EPCC (Edinburgh)

Option 2 in practice: Community CMS solution + EUDAT Safe Replication

DCH-RP <-> EUDAT PoC 2



Memory institutions:

EUDAT Simple Store

Provided by PSNC or EUDAT partners

EUDAT Safe replication

Provided by PSNC (Poznan) & EPCC (Edinburgh)

Domain-specific solutions dLibra/dArceo/dLab

DCH-RP <-> EUDAT PoC 2

http://dlab.psnc.pl/

dLab
DIGITISATION WORKFLOW MANAGEMENT

dArceo
LONG-TERM PRESERVATION SERVICES

dMuseion
DIGITAL MUSEUM FRAMEWORK

dLibra
DIGITAL LIBRARY FRAMEWORK

Basic statistics:

- ± 100 digital libraries
- several hundreds memory institutions
- over 1,1 M of digital objects
- 98% content delivered via services based on dLibra software (http://dlibra.psnc.pl/) which uses Solr for content and metadata indexing and searching

EUDAT <-> DCH-RP PoC summary

- **We investigate two possible ways to offering data preservation services for DCH**
  - Top-down solution for ‚citizen scientists' / ‚citizen DCH people'
  - Well-established solution backed by EUDAT Safe Replication
- **We exploit the layered EUDAT CDI architecture**
  - In theory: It enables integration with existing solutions
  - We try to understand how it works in practice

Planned extensions of simple store
Discussion trigger 1

Development roadmap

- Premium service:
  - Customisation for layout, metadata – for community needs
  - increased storage capcity
  - increased support
  - increased bandwidth

- Premium vs regular service:
  - Providing premium service requires enrolling with EUDAT
  - Regular services to be offered to ‚citizen scientists/users' – no close relationship needed

- AAI integration
  - *On the roadmap*

Possible extensions of simple store
Discussion trigger 2

- **From yesterdays discussion about Simple Store:**

  - **Thousand of files?**
  - **Upload reliability / robustness?**

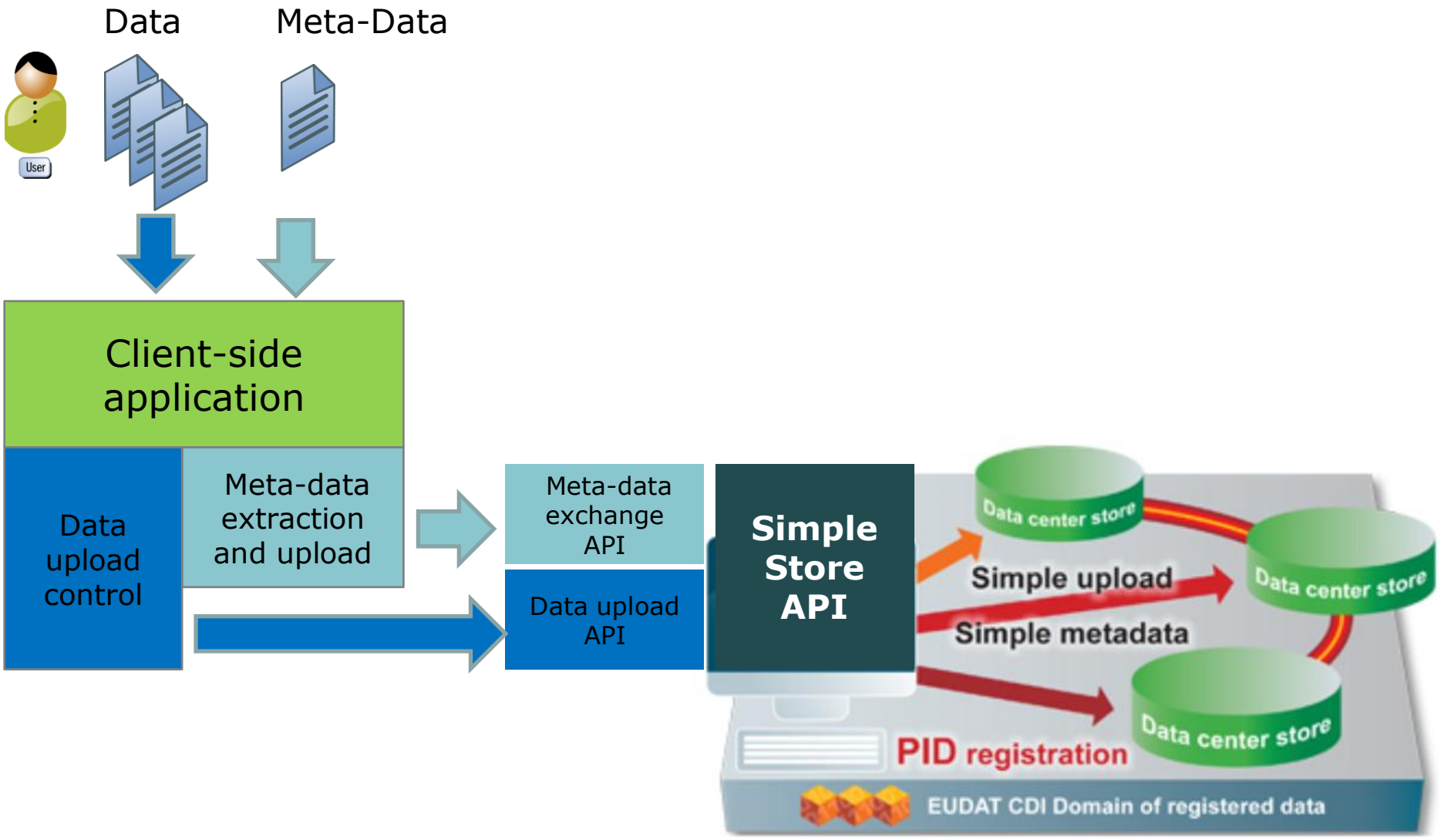
    => Batch upload
    - API to be developed
      - » Enables integration with existing systems
      - » Tools can be offered by EUDAT to support batch upload

  - **Collections upload?**

    => Support for meta-data extraction
    - Implemented client-side?
    - E.g. based on pre-prepared collections (e.g. DIPs)

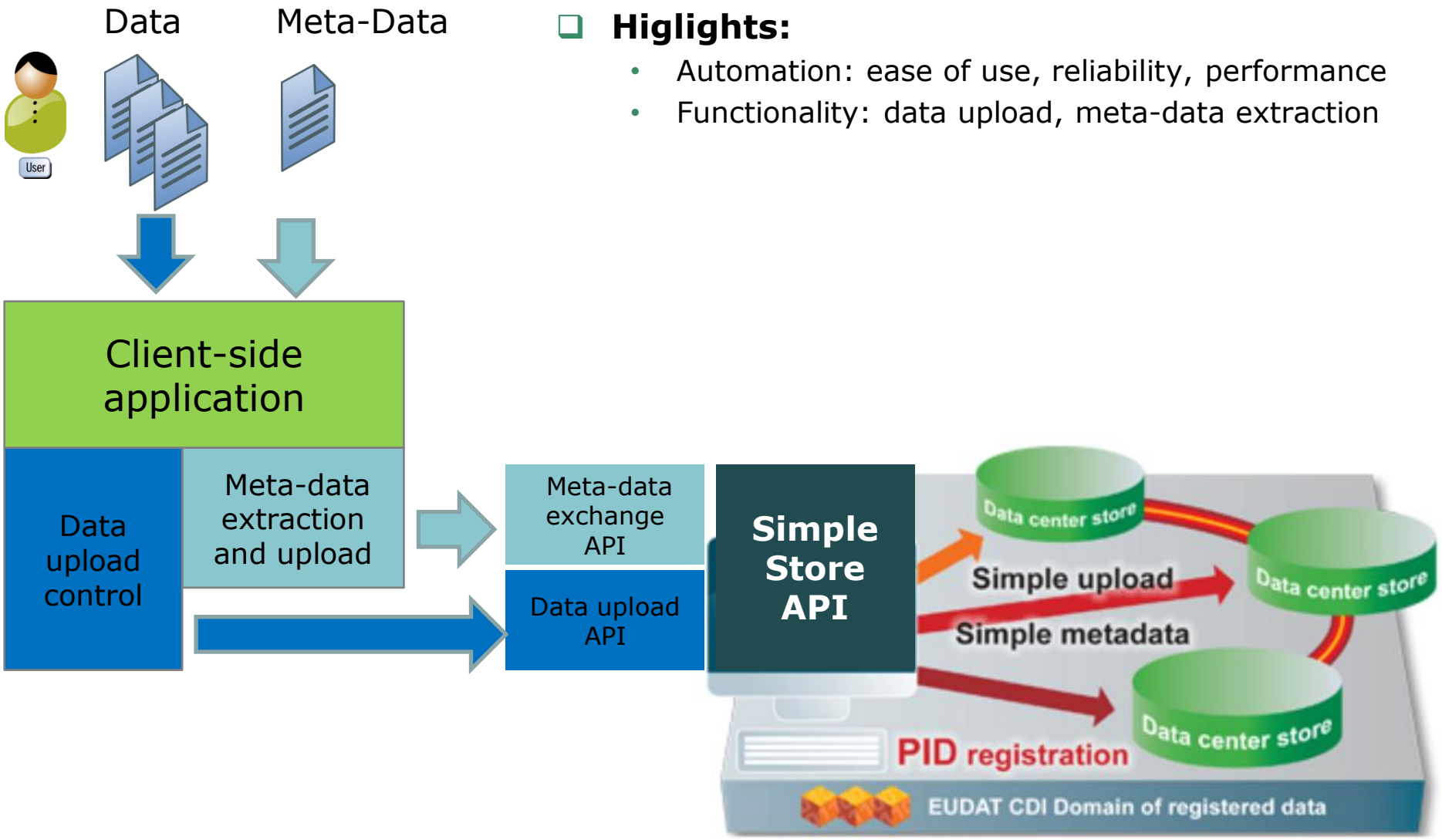
Possible extensions of simple store

Implementation: user-side tool?



Possible extensions of simple store

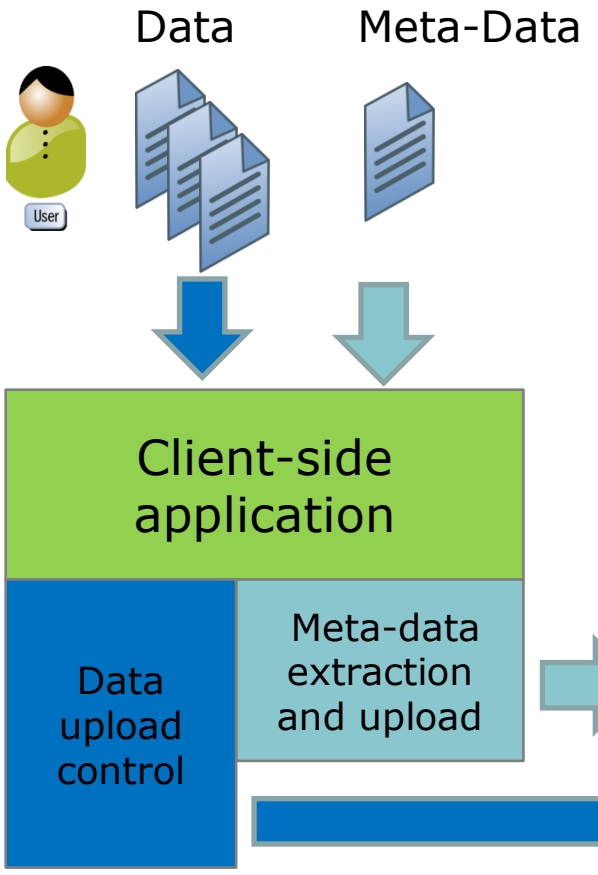
Implementation: user-side tool?



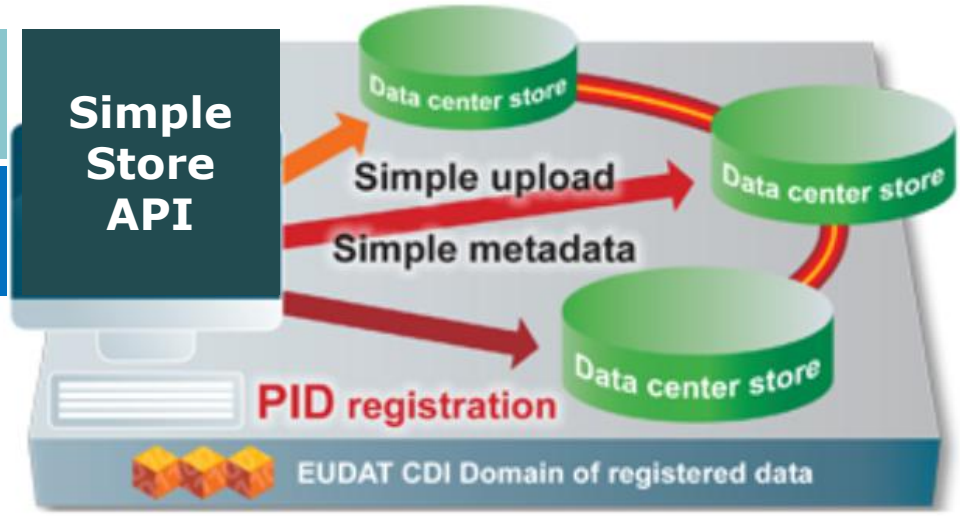
- **Higlights:**
  - Automation: ease of use, reliability, performance
  - Functionality: data upload, meta-data extraction

Possible extensions of simple store

Implementation: user-side tool?

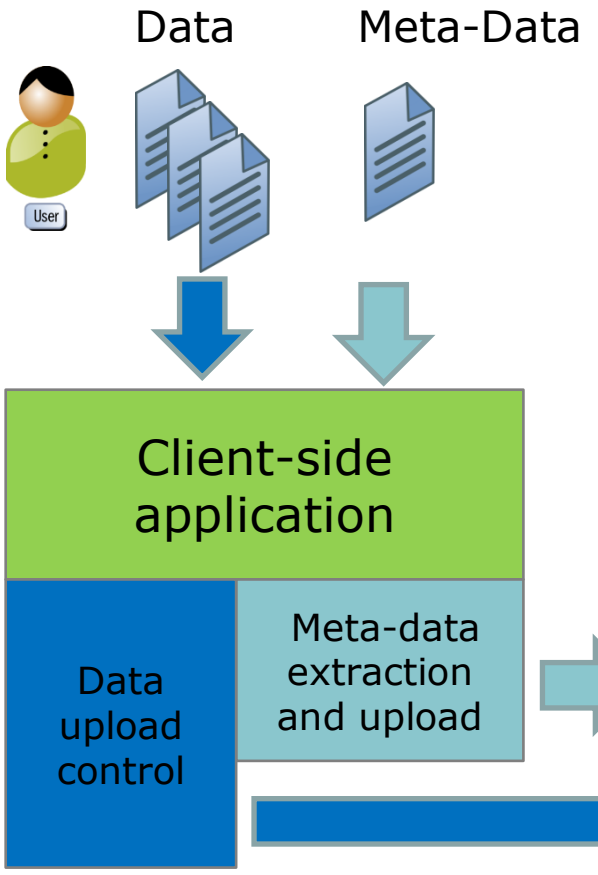


- **Higlights:**
  - Automation: ease of use, reliability, performance
  - Functionality: data upload, meta-data extraction
- **Challenges:**
  - Portability
  - Universality: standards need to be identified
  - Sustainability

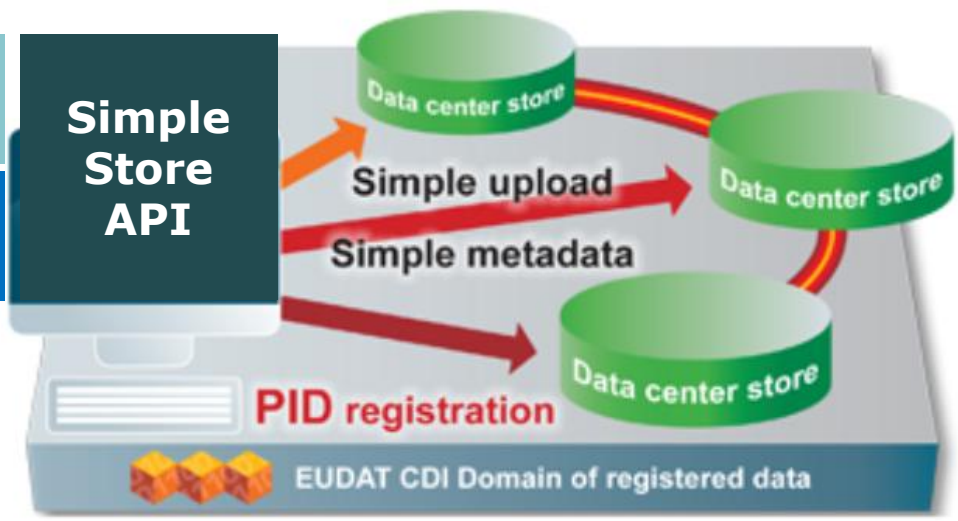


Possible extensions of simple store

Implementation: user-side tool?

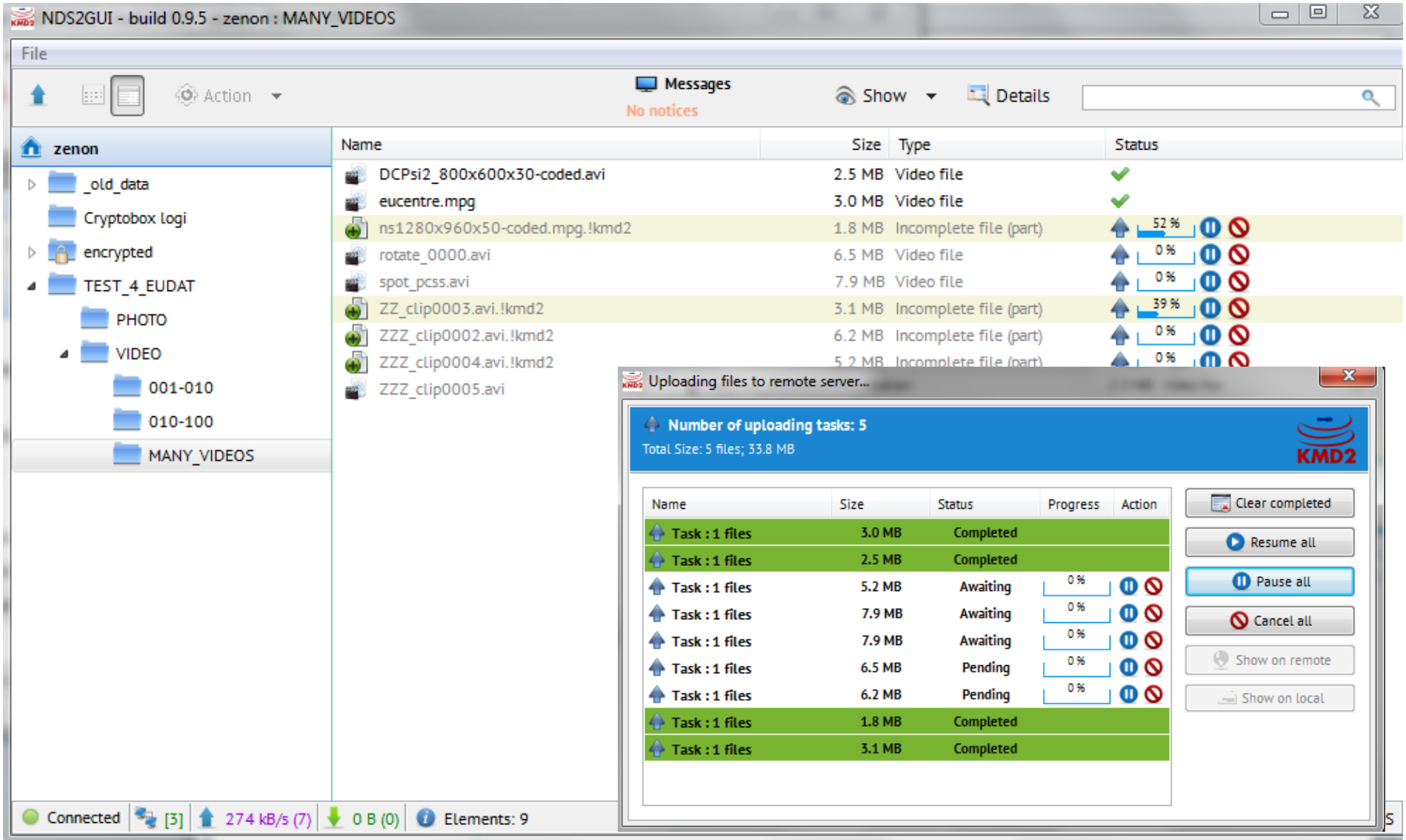


- **Higlights:**
  - Automation: ease of use, reliability, performance
  - Functionality: data upload, meta-data extraction
- **Challenges:**
  - Portability
  - Universality: standards need to be identified
  - Sustainability
- **Discussion needed!**



Possible extensions of simple store

Implementation: user-side tool?



Extensions of EUDAT services discussion
Summary

- **Message:**
  - EUDAT infrastructure and services are layered, modular
    - This enables integration
  - Further extensions possible
    - Users are welcome to influence them
  - We want to make sure that we recognised and support necessary standards
  - Technical details / organisation / etc. to be discussed

Using EUDAT services to replicate, store, share, and find cultural heritage data

…in PSNC… and beyond…

Maciej Brzeźniak, Norbert Meyer, PSNC
Damien Lecarpentier, CSC

THANK YOU!