European Data Infrastructure - EUDAT

Data Services & Tools

Dr. – Ing. Morris Riedel
Research Group Leader, Juelich Supercomputing Centre
Adjunct Associated Professor, University of iceland
BDEC2015, 2015-01-28

Relevance of Solving Big Data Challenges towards Exascale

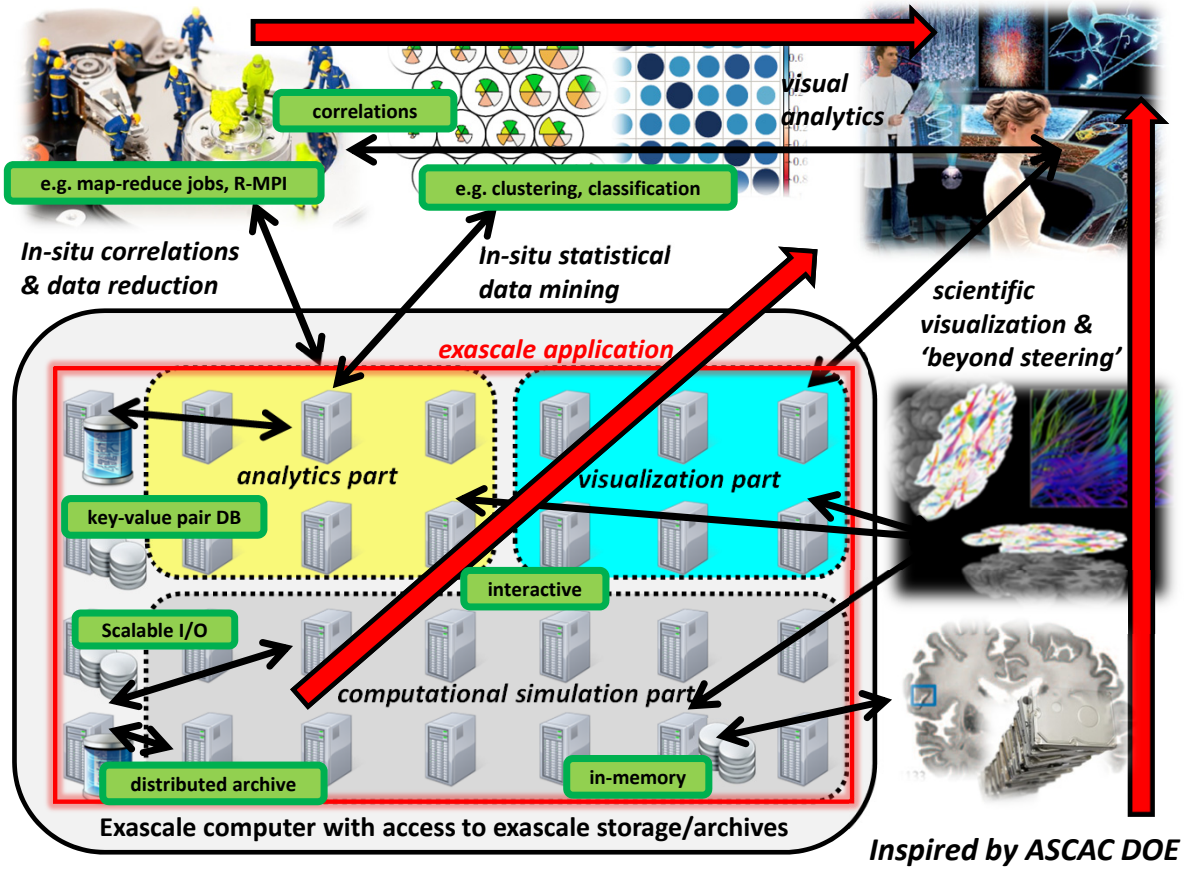
How to reduce?

How to re-use?

How to identify duplicates?

How to link?

How to access?



Inspired by ASCAC DOE report

How to find?

How to share?

How to replicate?

How to annotate?

How to transfer?

Motivation - ‘Need for Big Data Tools‘ in HPC & Exascale

Ever increasing volumes, varieties, velocities

- Shift from tape to active disks → active processing
- Data transfer-aware scheduling → transfer takes time
- Different copies of ‘same data‘ → sharing data necessary
- Different copies of ‘same data‘ in different epresentations → delete some data (e.g. tool-dependent data types, e.g. libsvm format vs. Original image, etc.)

Publication process changes

- Open referencable data is required for journals → data publicly available
- Long-lasting copies years after HPC users finished projects → archiving
- Technology changes, links need to persist in papers → handle systems

New toolsets

- Data replication, in-memory & data sharing tools, different filesystems, etc.
- Statistical data mining codes for classification, clustering, applied statistics, etc. (potential to validate, e.g. inverse problems, or reduce datasets, e.g. PCA)

EUDAT

EUDAT: A pan-European e-Infrastructure with useful tools

- Computational scientists are facing ‘big data’ challenges
  - Where to store the growing amount of data?
  - How to find it & how to reduce large quantities of data?
  - How to re-use & share or archive for publishing?

- Many communities are developing own solutions → Good
  - … but solutions need to remain interoperable and sustainable for centers

- EUDAT offers a pan-European solution
  - Providing a set of generic tools to help managing growing amount of data
  - Providing tools across communities to ensure minimum level of interoperability
  - Linking community specific repositories to the largest European scientific data and HPC centers → Collaborative Data Infrastructure (CDI)

EUDAT

Data Centers and Communities

26 European partners

EBERHARD KARLS UNIVERSITÄT TÜBINGEN
KIT Karlsruhe Institute of Technology
INGV
CERN
epcc
BSC Barcelona Supercomputing Center Centro Nacional de Supercomputación
DKRZ DEUTSCHES KLIMARECHENZENTRUM
CINES Centre Informatique National de l'Enseignement Supérieur
CINECA
RedIRIS
CSC
rzg RECHEN-ZENTRUM GARCHING
Max-Planck-Institut für Meteorologie
JÜLICH FORSCHUNGSZENTRUM
Science & Technology Facilities Council
maatG
CERFACS
umweltbundesamt ENVIRONMENT AGENCY AUSTRIA
POZNAN
SNIC
Trust-IT Services Ltd Communicating ICT to markets
UNINETT sigma
SURF SARA
UCL

EUDAT

User Forums + 30 communities

GBIF

1st User Forum
7-8 March 2012, Barcelona

CLARIN
Common Language Resources and Technology Infrastructure

enes
EUROPEAN NETWORK FOR EARTH SYSTEM MODELLING

LifeWatch

2nd User Forum
11-12 March 2013 London

BioVeL

Virtual Physiological Human
network of excellence

EPOS
EUROPEANPLATEOBSERVINGSYSTEM

incf International Neuroinformatics Coordinating Facility

VERCE

european multidisciplinary seafloor observatory emso

BBMRI
Biobanking and Biomolecular Resources Research Infrastructure

EIDA
European Integrated Data Archive

DRIHM
DISTRIBUTED RESEARCH INFRASTRUCTURE FOR HYDRO-METEOROLOGY

DCH RP

FOT-NET

neuGRID

DARIAH
Digital Research Infrastructure for the Arts and Humanities

EISCAT

dixa

Geo-Seas

WLCG
Worldwide LHC Computing Grid

SCIDIP-ES
SCIENCE DATA INFRASTRUCTURE FOR PRESERVATION - EARTH SCIENCE

pandata europe

marine

LTER europe

icos integrated carbon observation system

cessda
Council of European Social Science Data Archives

Science & Technology Facilities Council
ISIS

elixir

MAPPER

IAGOS

ECRIN
EUROPEAN CLINICAL RESEARCH INFRASTRUCTURES NETWORK

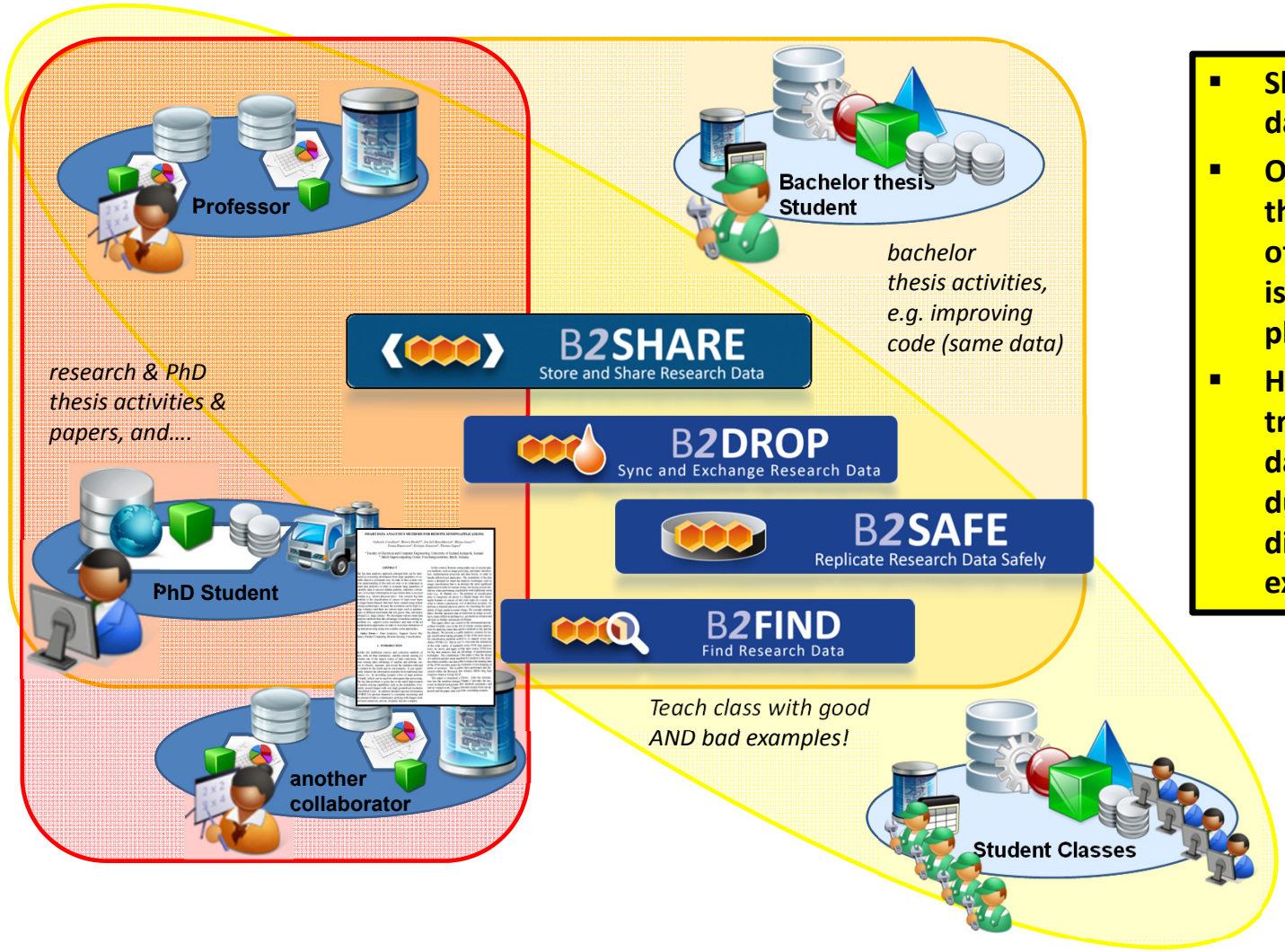
aginfra
A data infrastructure for agriculture

EURO VO AIDA Astronomical Infrastructure for Data Access

EMBRC
EUROPEAN MARINE BIOLOGICAL RESOURCE CENTRE

EUDAT

Work situation in scientific computing…

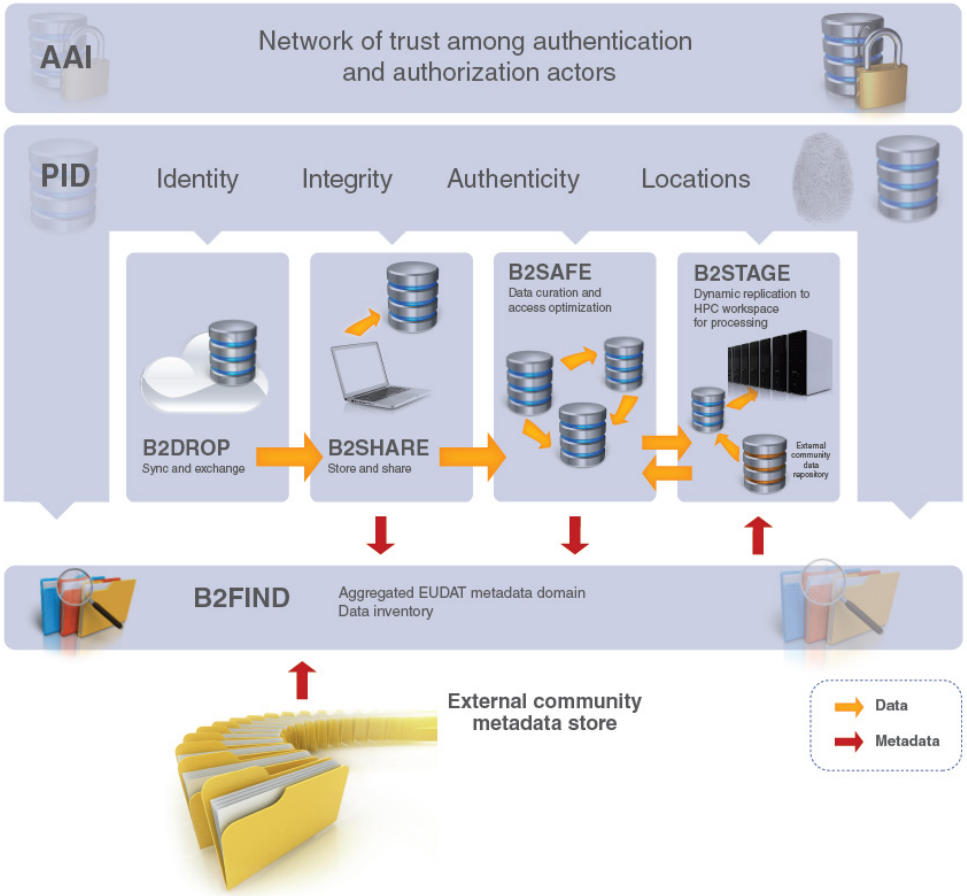


- **Simple tools are important**
- **Avoid overheads in data management**
- **Realistic use within HPC environments**

- **Sharing different datasets is key**
- **One tend to loose the overview of which data is stored on which platform**
- **How do we gain trust to delete data when duplicates on different systems exist**

EUDAT

Toolset Overview



- **Access** and **deposit**,
- **Informal data sharing**
- **Long-term archiving**,
- Addressing **identification**, **discoverability** and **computability**
- **long-tail and 'big' data**

→ **address full lifecycle of research data**

→ **adopt only what is needed**

B2DROP

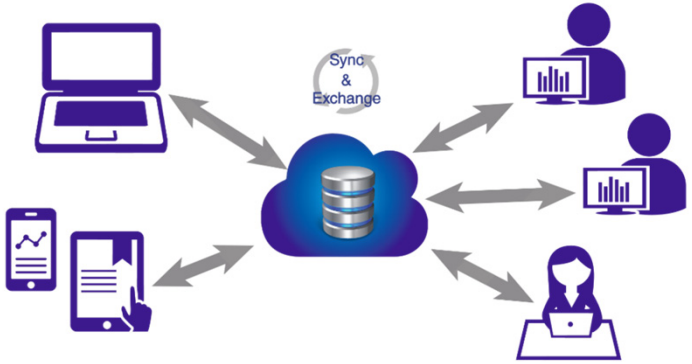
Sync and Exchange Research Data

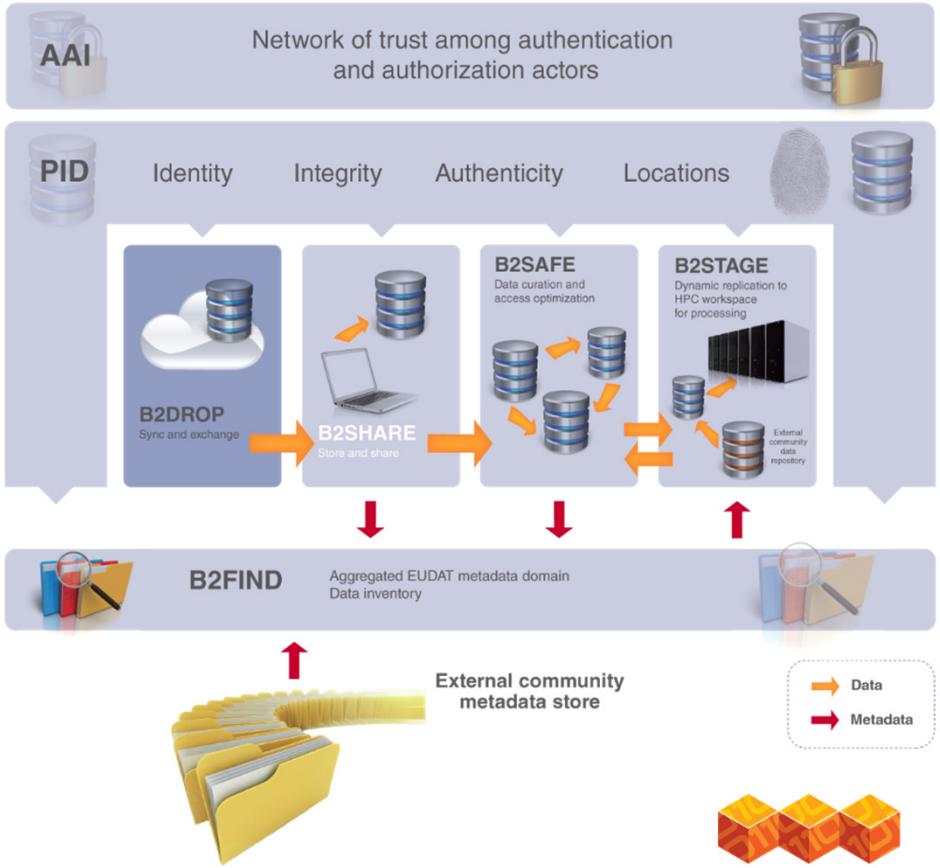
b2drop.eudat.eu

B2DROP is a **secure and trusted** data exchange service for researchers and scientists to keep their research data **synchronized** and up-to-date and to **exchange** with other researchers.

An ideal solution to:

- **Store and exchange** data with colleagues and team
- **Synchronize** multiple versions of data
- Ensure **automatic desktop** synchronization of large files





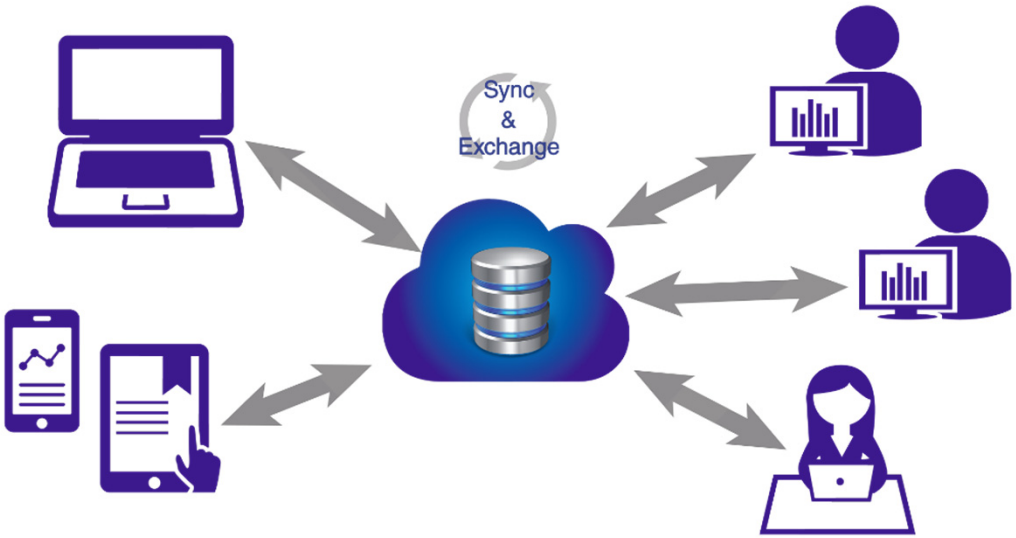
EUDAT

Features

B2DROP
Sync and Exchange Research Data

b2drop.eudat.eu

- future **integration with the B2 suite of services** to allow user-friendly data sharing
- users decide **with whom to exchange data, for how long and how**
- **up to 20GB of storage space** for research data
- **access and manage permissions** to files **from any device and any location**
- **simple to use and open to all** researchers, scientists, communities alike to **synchronize and exchange data with one or multiple users**



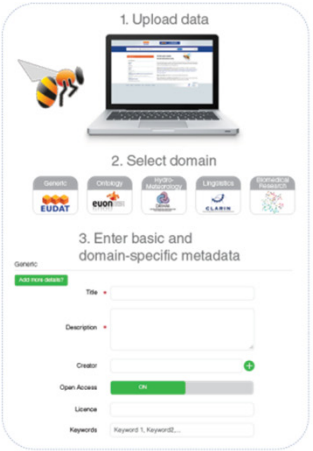
B2SHARE is a **user-friendly**, **reliable** and **trustworthy** way for researchers, scientific communities and citizen scientists to **store** and **share** small-scale research data from diverse contexts.

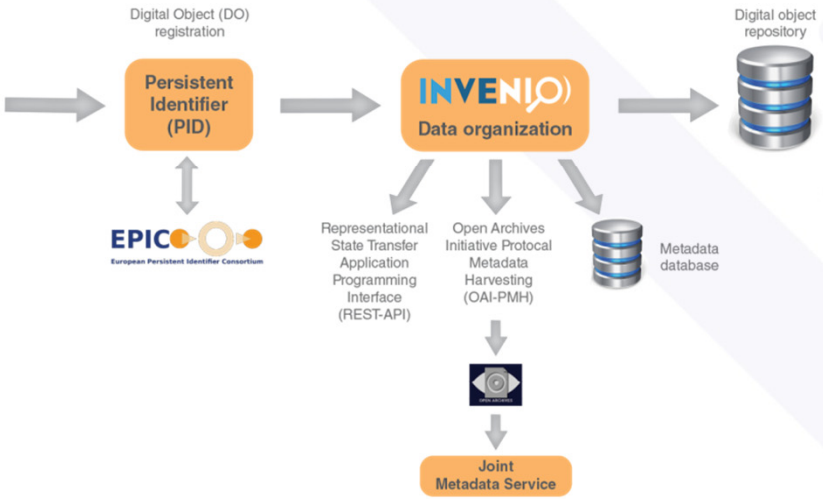
B2SHARE
Store and Share Research Data

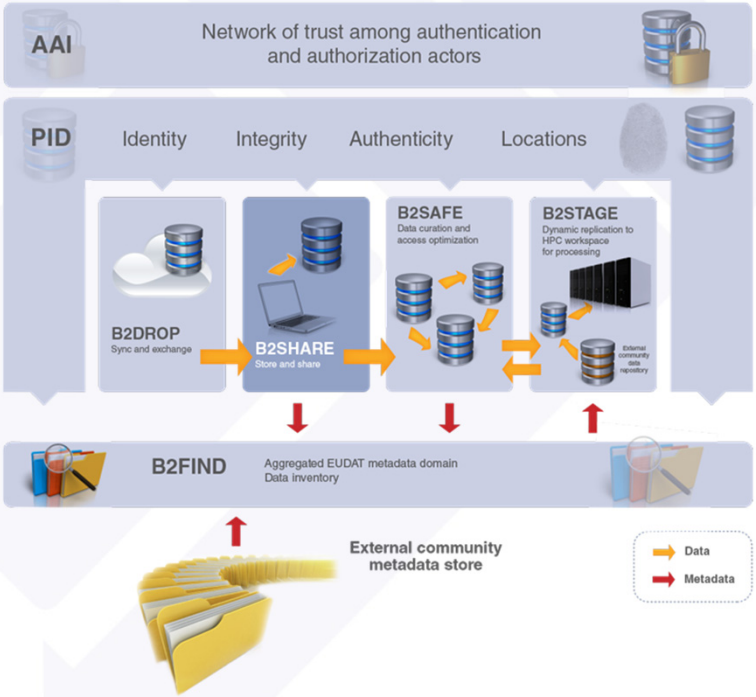
b2share.eudat.eu

A winning solution to:

- **Store:** facilitates research data storage
- **Preserve:** guarantees long-term persistence of data
- **Share:** allows data, results or ideas to be shared worldwide







EUDAT

Features

B2SHARE
Store and Share Research Data

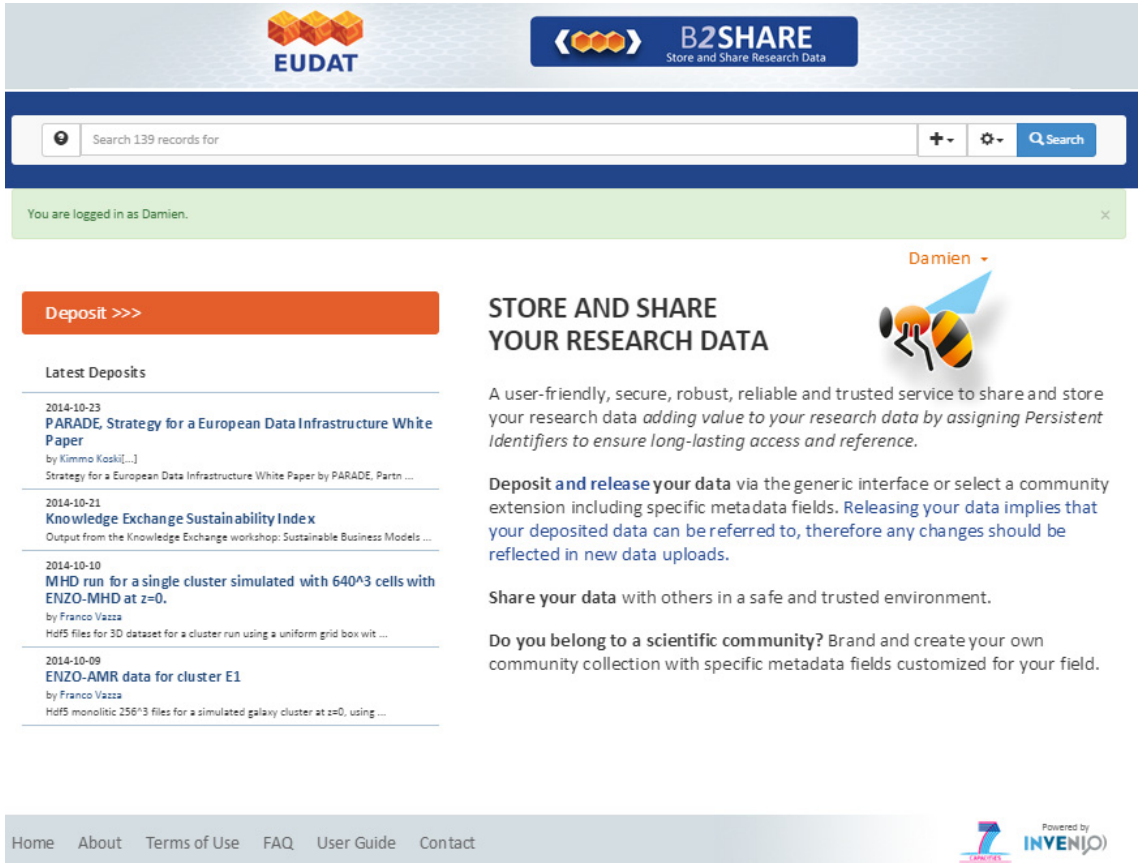
b2share.eudat.eu

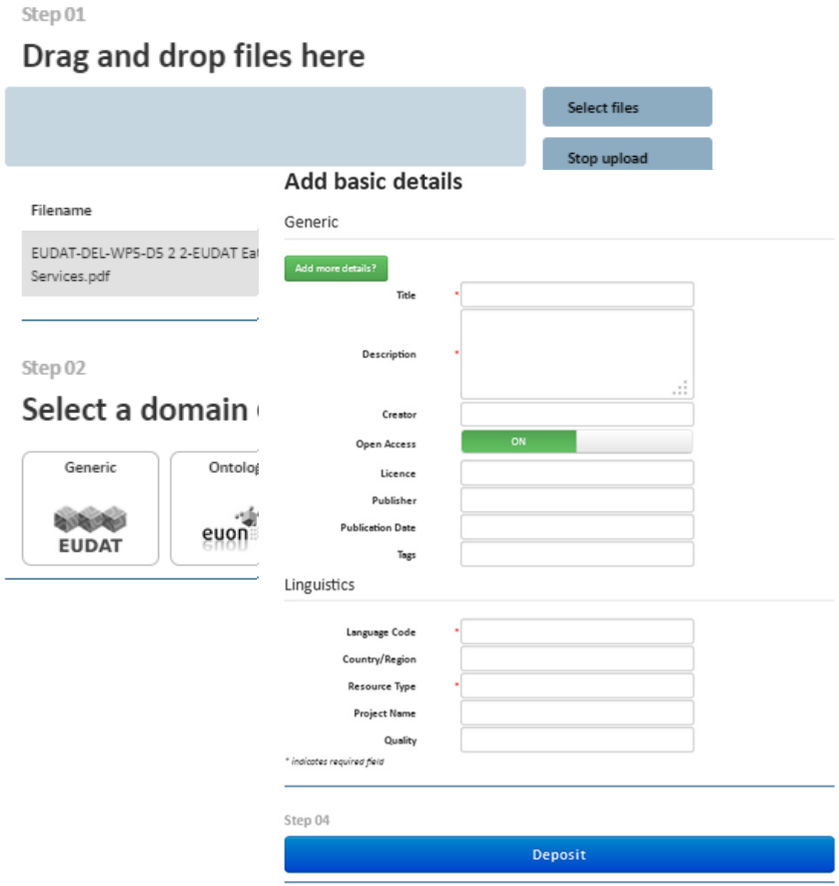
- Targets **small-scale research data collected as part of international collaboration and looking for a central repository**
- integrated with the **EUDAT collaborative data infrastructure**
- **free upload and registration** of stable research data
- data assigned a **permanent identifier**, which can be retraced to the data owner
- **community-specific metadata extensions** and **user interfaces**
- **openly accessible** and **harvestable metadata**
- **representational state transfer application programming interface (REST API)** for integration with community sites
- **data integrity ensured by checksum** during data ingest
- **professionally managed storage service –** no need to worry about hardware or network
- **monitoring of availability and use**

‘A four-click service’

B2SHARE
Store and Share Research Data

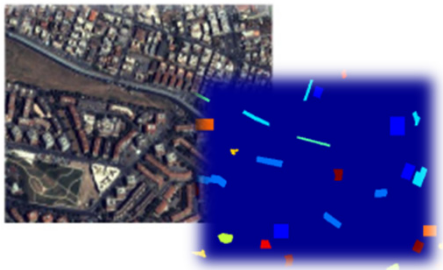
b2share.eudat.eu



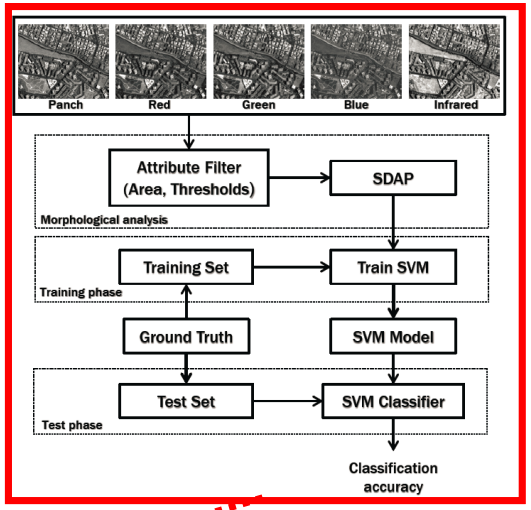


HPC Usage Example

Classification
(one field in data mining)



Sattelite Data (Quickbird)



Parallel
Support Vector
Machines (SVM)

Classification Study of Land Cover Types

HPC / MPI code

B2SHARE
Store and Share Research Data

b2share.eudat.eu

Class	Training	Test
Buildings	18126	163129
Blocks	10982	98834
Roads	16353	147176
Light Train	1606	14454
Vegetation	6962	62655
Trees	9088	81792
Bare Soil	8127	73144
Soil	1506	13551
Tower	4792	43124
Total	77542	697859

```
#!/bin/bash
#MSUB -N Train-tune-rec86-4-16-32
#MSUB -l nodes=4:ppn=16:performance
#MSUB -l walltime=03:00:00
#MSUB -M m.riedel@fz-juelich.de
#MSUB -m abe
#MSUB -W x=naccesspolicy:singlejob
#MSUB -v tpt=2
#MSUB -q devel


### jobscript

cd $PBS_O_WORKDIR
echo "workdir: $PBS_O_WORKDIR"

NSLOTS=32

echo "running on $NSLOTS cpus..."

### location
PISVM=/homeb/zam/mriedel/pisvm-1.2/pisvm-1.2/pisvm-train

TRAINDATA=/homeb/zam/mriedel/bigdata/86-
romeok/sdap_area_all_training.el

### submit
mpiexec -np $NSLOTS $PISVM -o 1024 -q 512 -c 10000 -g 16 -t 2 -m
1024 -s 0 $TRAINDATA
```

„Reference Data Analytics“
for reusability & learning

CRISP-DM Report

Openly Shared Datasets

Running Analytics Code

- **HPC JobScripts**
- **HPC run in-/outputs**
- **Input data & metadata**
- **Output data & metadata**
- **PIDs for Trust to Delete**
- **Handle for Publications**

Search for PiSVM Big Data Analytics in B2SHARE

B2SAFE is a **robust, safe and highly available service** which allows community and departmental repositories to **implement data management policies on their research data** across multiple administrative domains in a trustworthy manner.

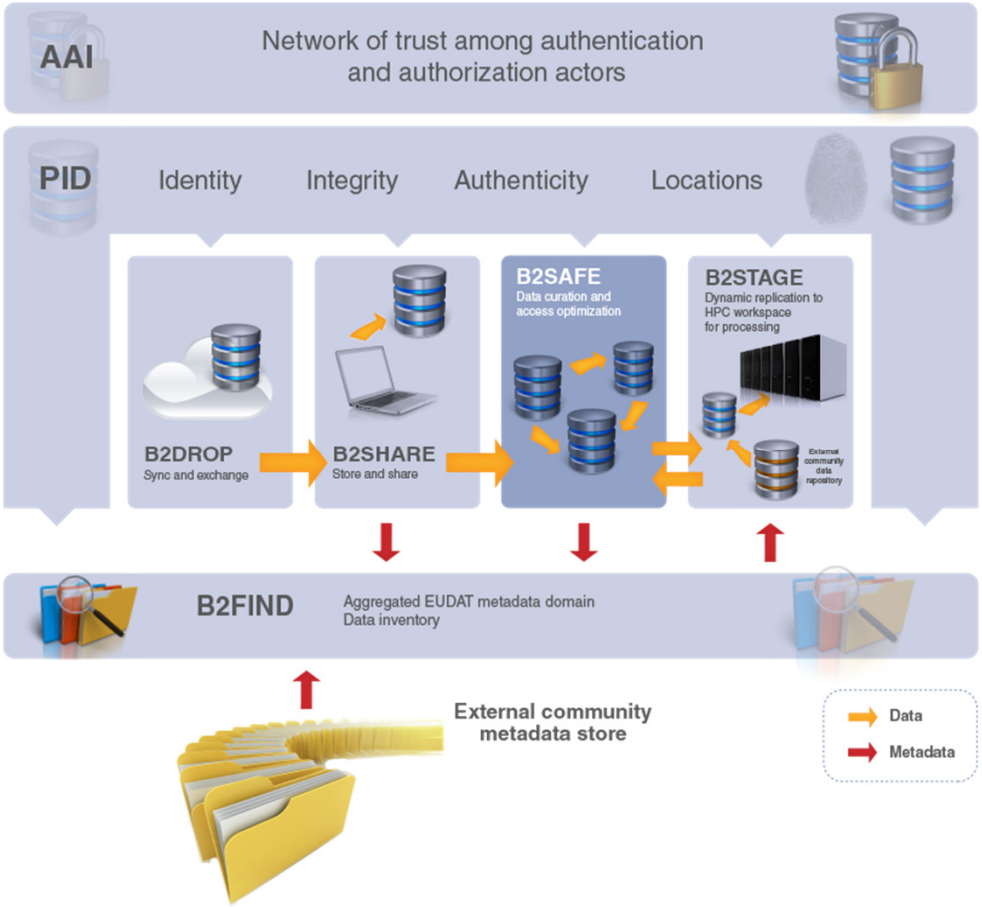
A solution to:

- 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** for compute-intensive analysis

B2SAFE

Replicate Research Data Safely

eudat.eu/b2safe

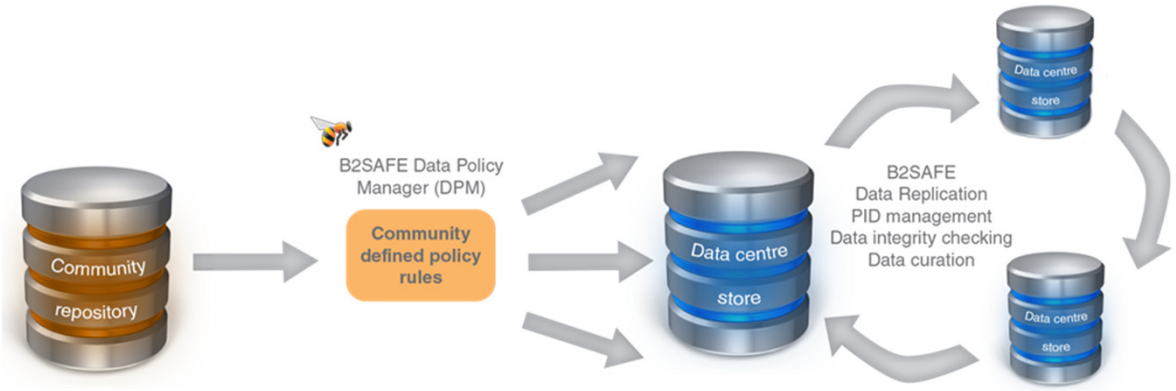


Features

B2SAFE
Replicate Research Data Safely

eudat.eu/b2safe

- based on the execution of **auditable data policy rules** and the use of **persistent identifiers (PIDs)**
- respects the **rights of the data owners** to define the **access rights for their data** and to decide how and when it is made **publicly referenceable**



- data policies are **centrally managed via a Data Policy Manager**, and the policy rules are implemented and enforced by **site-local rule engines**
- able to **aggregate data from different disciplines** into a storage system of **trustworthy and capable data service providers**
- support for **repository packages** (e.g. DSPACE, FEDORA) and a **lightweight HTTP-based solution**

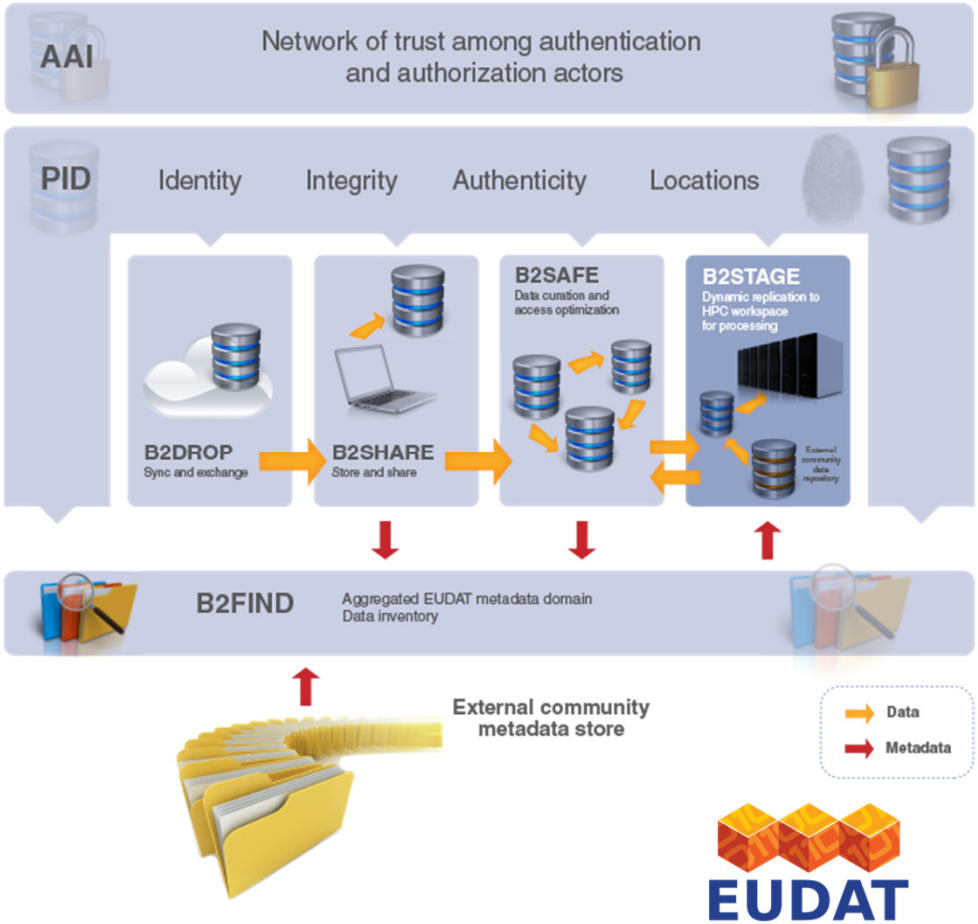
B2STAGE
Get Data to Computation

eudat.eu/b2stage

B2STAGE is a **reliable, efficient, light-weight and easy-to-use** service to **transfer research data sets** between EUDAT storage resources and high-performance computing (HPC) workspaces.

The service allows users to:

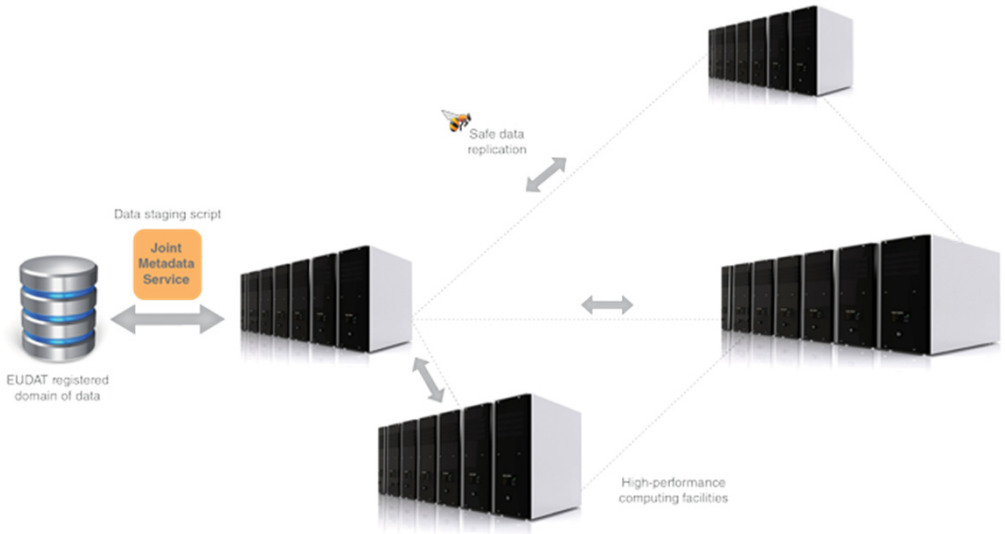
- **Transfer** large data collections from EUDAT storage facilities to **external HPC facilities for processing**
- In conjunction with B2SAFE, **replicate community data sets, ingesting them** onto EUDAT storage resources for long-term preservation
- **Ingest computation results** into the EUDAT infrastructure
- Access data through a **RESTful HTTP interface** (in progress)



EUDAT

Features

- **an extension of the B2SAFE and B2FIND services,** which allow users to store, preserve and find data
- **data-staging script** facilitates staging, ingestion and retrieval of persistent identifier (PID) information of transferred data
- service available **to all registered researchers and interested communities**



B2STAGE
Get Data to Computation

eudat.eu/b2stage

- **users negotiate access to remote HPC services in parallel**
- **collaboration with other infrastructures,** such as the European Grid Infrastructure (EGI) and Partnership for Advanced Computing in Europe (PRACE)
- **documentation, educational material and service helpdesk** available to support users

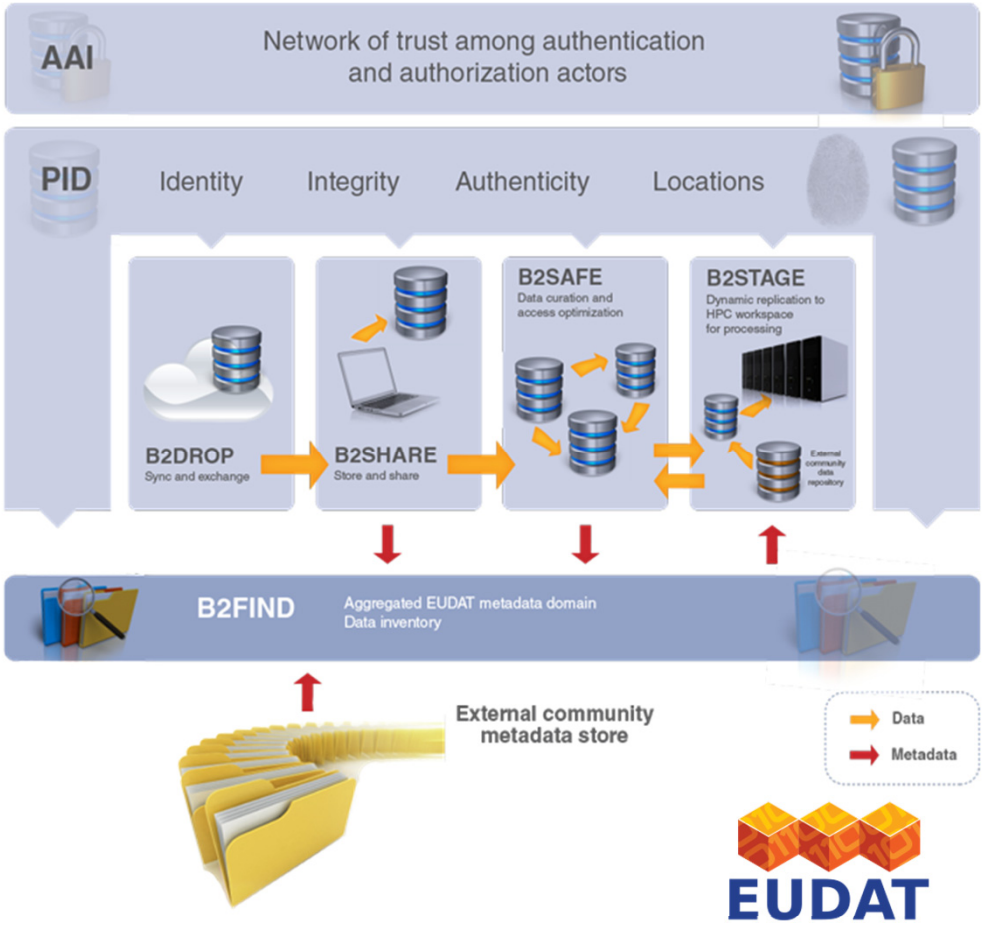
B2FIND is a simple, user-friendly **metadata catalogue of research data collections** stored in EUDAT data centres and other repositories.

A service which allows users to:

- **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**

B2FIND
Find Research Data

b2find.eudat.eu

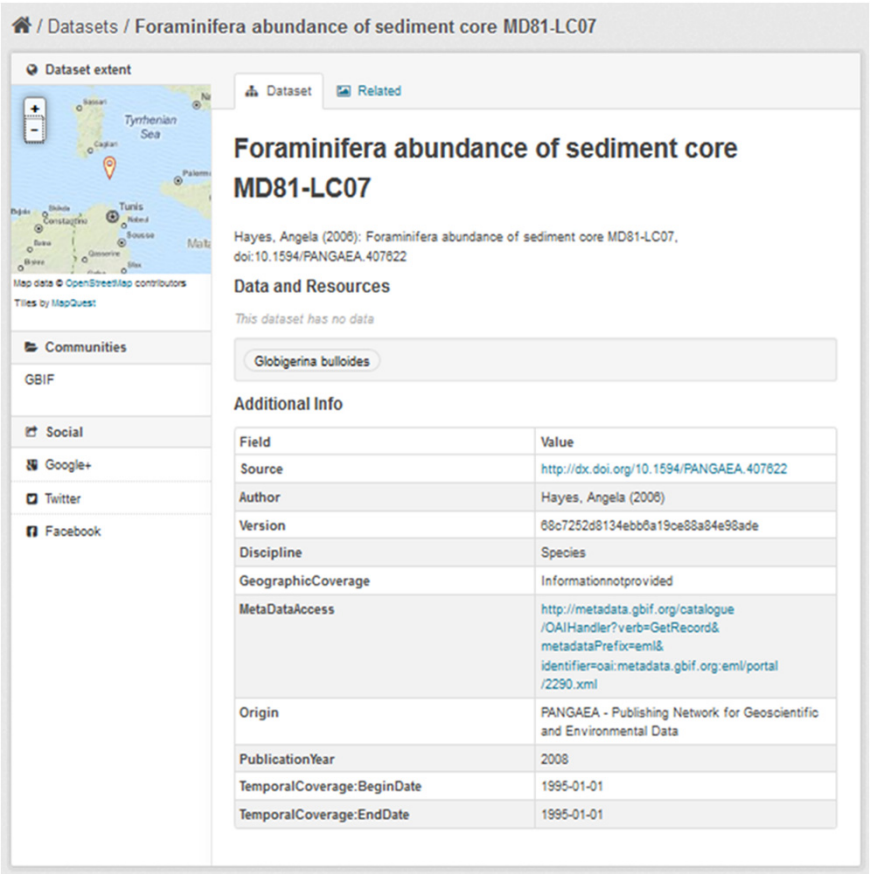


Features

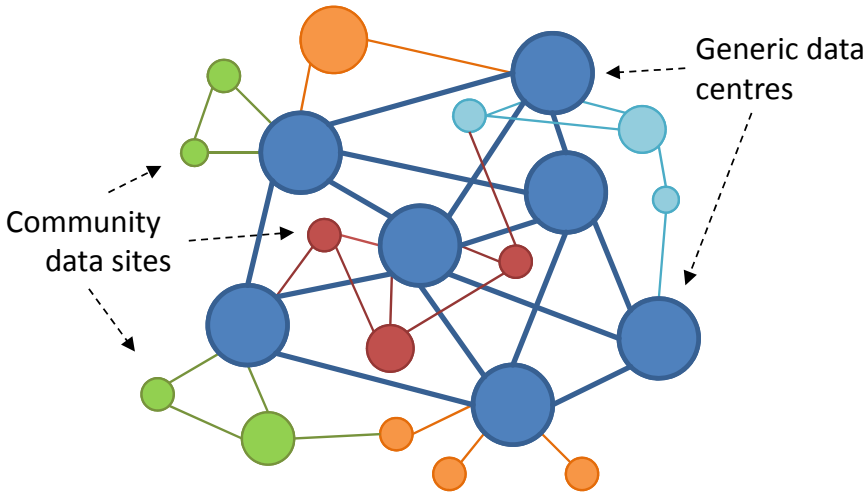
B2FIND
Find Research Data

b2find.eudat.eu

- **supports faceted, geospatial and temporal metadata searches**
- allows users **to search and browse datasets via keyword searches**
- initially available for **communities in the EUDAT registered domain of data**
- EUDAT will then extend the service **to other interested and reliable data and metadata providers**
- results displayed in **user-friendly format** and **listed in order of relevance**
- access to **the scientific data objects** is given through **references provided in the metadata**



A Federated and Distributed CDI

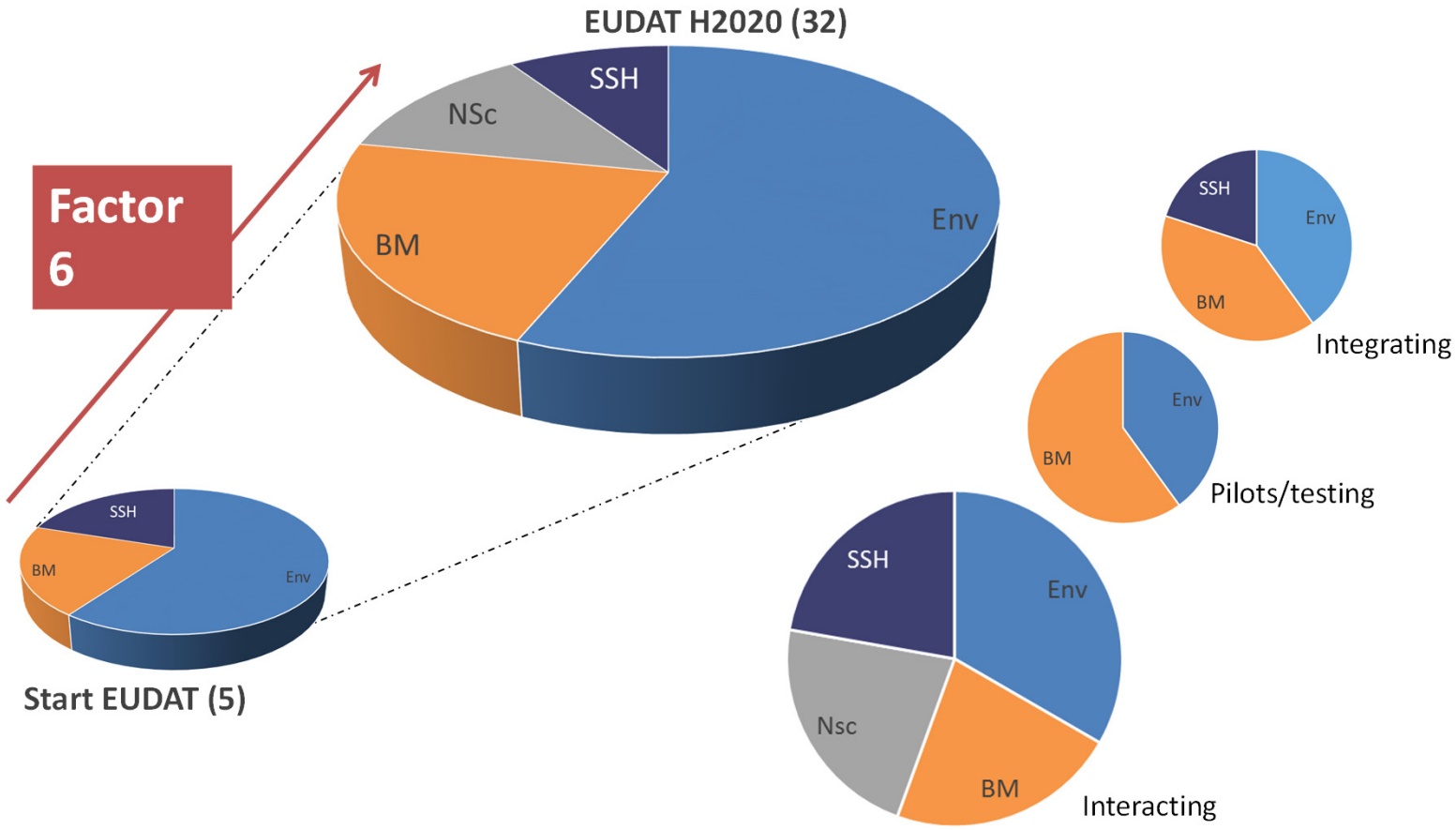


- ***Using EUDAT services:*** finding and accessing data, for instance, or storing smaller data sets by interacting with one of the CDI public front-end services

vs

- ***Joining the CDI:*** implies a tighter integration with at least one of the EUDAT centre → partnership between legal entities relying on OLAs and SLAs

Community Outreach & Service take Up



EUDAT Policies / Data Access and Reuse

- Open Access?
  - Funders: “Yes, absolutely!”
  - Researchers: “Yes, but…”
    - Some data is “sensitive”
    - What about credit and merit – others ‘harvesting’?
    - How to find one’s way in the legal minefield?
  - ‘Data-driven application-enabling’ activities
    - Providing tools and services to handle sensitive data
    - Licensing guidance, PIDs and usage statistics
    - Training & working on case studies (e.g. HPC simulation data demands)

OPEN DATA – WHAT DO EUDAT COMMUNITIES REALLY THINK ABOUT IT?

MARIE SANDBERG, PAWEŁ KAMOCKI, DAMIEN LECARPENTIER, ROB BAXTER

TO APPEAR IN ERCIM NEWS NO. 100 (JANUARY 2015) [1]

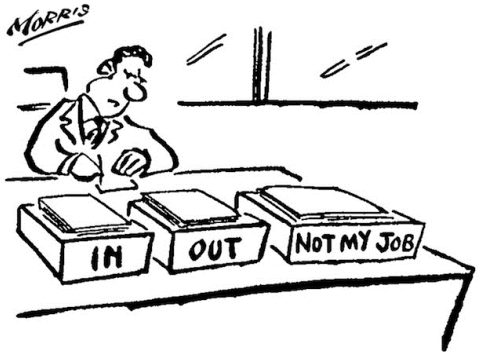
Facilitating open access to research data is a principle endorsed by an increasing number of countries and international organizations, and one of the priorities in the European Commission’s Horizon 2020 funding framework [2][3]. But what do researchers themselves think about it? How do they perceive the increasing demand for open access and what are they doing about it? What problems do they face, and what sort of help are they looking for?

As a pan-European research data infrastructure, these are questions that are of fundamental interest to EUDAT. To understand better what researchers think, EUDAT has conducted a programme of interviews with fourteen major research communities from the fields of life sciences, Earth and atmospheric science, astrophysics, climate science, biodiversity, agricultural science, social science and humanities – a broad cross-section of European research interests. While one cannot, of course, interpret the views of any given individual as the official position of a whole research community, they nevertheless provide useful information on the general attitude, requirements and challenges researchers face with regard to opening up their research data. In this article we report on our initial conclusions from this survey.

GROWING AWARENESS

Open access to research data is increasingly seen as a compelling principle in many research communities. There is a growing awareness of the global move towards open access, the potential benefits of it, and the necessity to implement open access policies within their disciplines. According to preliminary figures on the first wave of open data pilot projects in Horizon 2020, the opt-out rate among proposals submitted to the “open by default” categories was below 30%, and the opt-in rate among other proposals was around about the same. This underlines our findings in EUDAT– researchers are pretty happy about sharing their data.

CHALLENGES AHEAD



Need to Understand Computational Scientists

- Research Infrastructures → CDI users, partners & stakeholders
  - Uptake plans: work with computational scientists & HPC users to understand where data services make a difference

- It is not only about developing technical solutions, but also about defining the right partnership model
  - Take into account existing arrangements within pan-European research communities (organisational structure, funding schemes, business models, etc.)

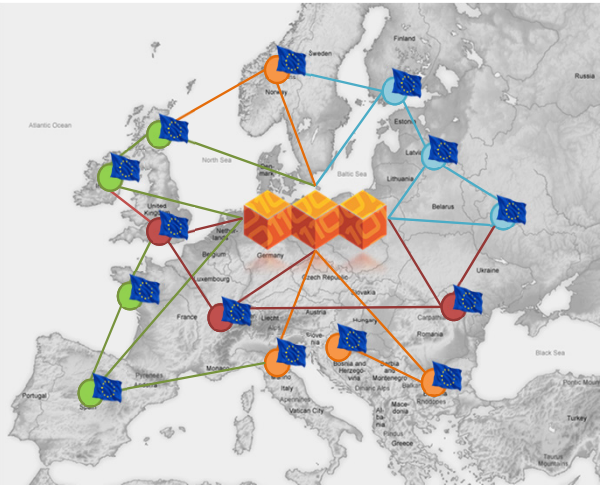
EUDAT

E-Infrastructure Commons

- Users have a “right” to a seamless access to network, data, and computing resources funded by public money
  - It is our role to make it as easy as possible for users →Users should not care about which e-Infrastructures they are using
- Cross-Infrastructure services
  - Based on pilots with interested communities
- E-Infrastructure Commons Roadmap

Bridging National and European Data Solutions

- Making national resources more available
  - Making visible valuable national collections through EUDAT
  - Access to European resources through national catalogues



- Enhancing cross-national collaborations
  - EUDAT provides a European extension to national solutions
  - True data sharing & archiving are pan-European challenges

EUDAT

Thank you

Talk available on
http://www.morrisriedel.de/talks

EUDAT