DATA REPOSITORIES

2nd Data Analysis Methods (DAMe) Workshop

14.03.2019 | Sander Apweiler | JSC

1. Data repository

2. EUDAT

2.1 overview

2.2 service suite

3. Hands-on

Data repository

WHAT IS A DATA REPOSITORY?

A **data repository** is a software suite to **store** and aggregate **data sets** from multiple sources. The data can be stored unstructured **with additional metadata**, describing the data itself. The access to the data can be limited, if needed, before the data is reused by other researchers.

$\Rightarrow$ A data repository is a collection of (domain specific) data sets, and additional metadata, for secondary use.

WHAT IS A DATA REPOSITORY NOT?

- An extension for your hard drive.
- A storage for volatile data.
- A storage for sensitive/personal (raw) data.

POSSIBLE CONSTRAINTS

- The data repository is public/private.
- The access to the data must be restricted.
- The data has an embargo date.
- Data protection issues in data.
- The storage is subject to legal restrictions, like storage for medical data.
- The availability is not guaranteed to the retention period of the data.

REASONS FOR USING

- A lot of initiatives/projects/guidelines/... which promote „open data“ and shared data like:
  - (GO)FAIR
  - RDA
  - „Gute wissenschaftliche Praxis“
  - Funding agencies like BMBF, EU Commission
  - Publications
- You want to share your results and data with other researchers.
- They support you to hold the retention period.

FAIR PRINCIPLES

- Findable
- Accessible
- Interoperable
- Reuseable

FAIR PRINCIPLES

FAIR[1] principles:

- 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

[1] https://www.doi.org/10.1038/sdata.2016.18

FAIR PRINCIPLES

FAIR[1] principles:

- 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

[1] `https://www.doi.org/10.1038/sdata.2016.18`

FAIR PRINCIPLES

FAIR[2] principles:

- 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

[2] https://www.doi.org/10.1038/sdata.2016.18

FAIR PRINCIPLES

FAIR[2] principles:

- 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
- Reuseable
  - 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

[2] https://www.doi.org/10.1038/sdata.2016.18

DATA REPOSITORIES

- Arctic Data Center - For NSF[3] Polar Programs
- NOMAD CoE - Materials science
- PANGAEA - Earth & environmental science
- Publisher
  - Springer Nature
  - dataplanet
- Community based solutions
  - LTER Network Data Portal
- Research center based solutions
  - Zenodo

[3] National Science Foundation

LOOKUP SERVICES

- CESSDA - Consortium of European Social Science Data Archives
- DataCite
- DataONE - Data Observation Network for Earth
- re3data - Registry of research data repositories

EUDAT: Overview

EUDAT

EUDAT is a pan-European e-infrastructure solution for pan-European research infrastructure data challenges.

- All research infrastructures are facing data challenges
  - Where to store the growing amount of data?
  - How to find the data?
  - How to make the most of the data?

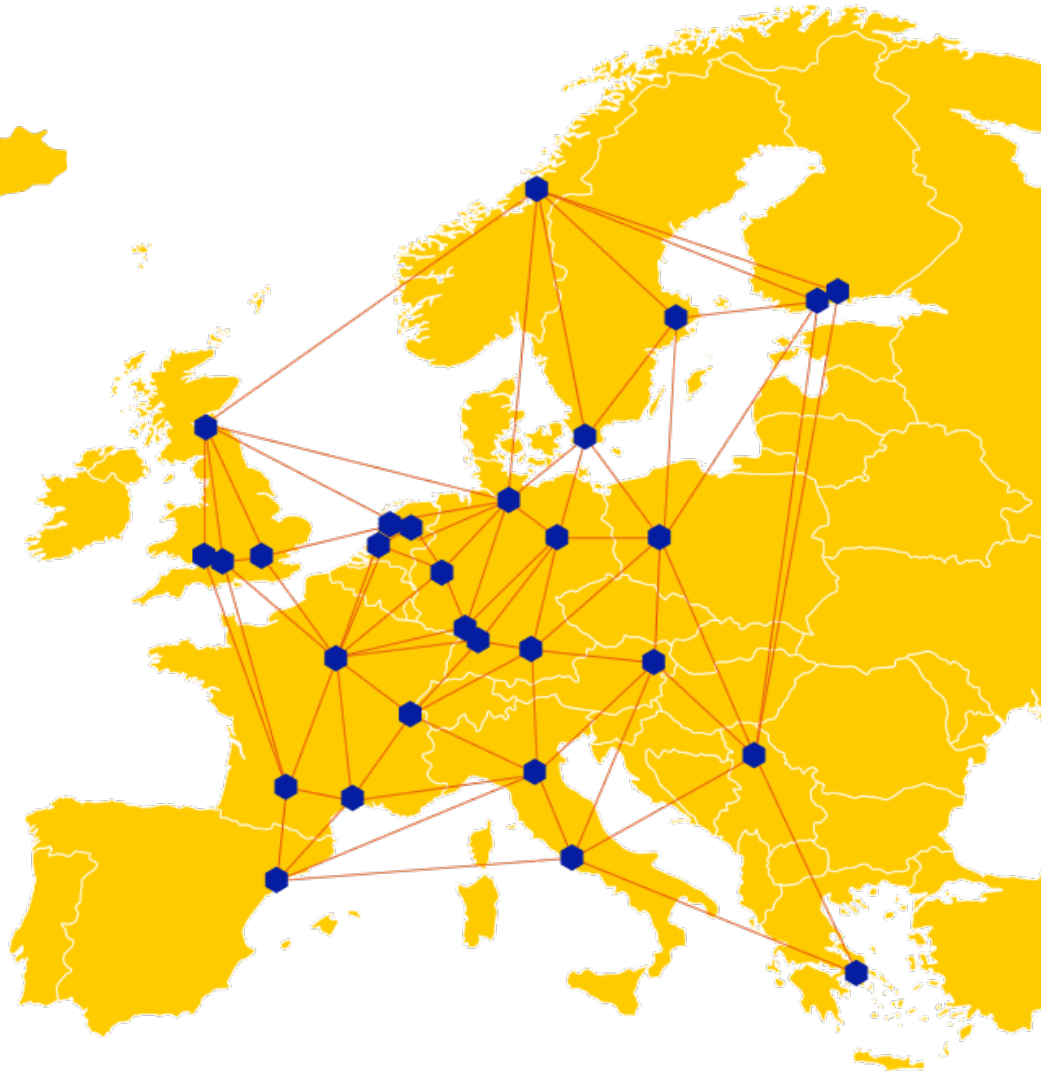
$\Rightarrow$ Solutions are needed at pan-European level.

- We need to promote synergies
  - Some services are common to many communities
  - Costs and investments can be optimised

$\Rightarrow$ Better integration of e-infrastructures and research infrastructures can be achieved.

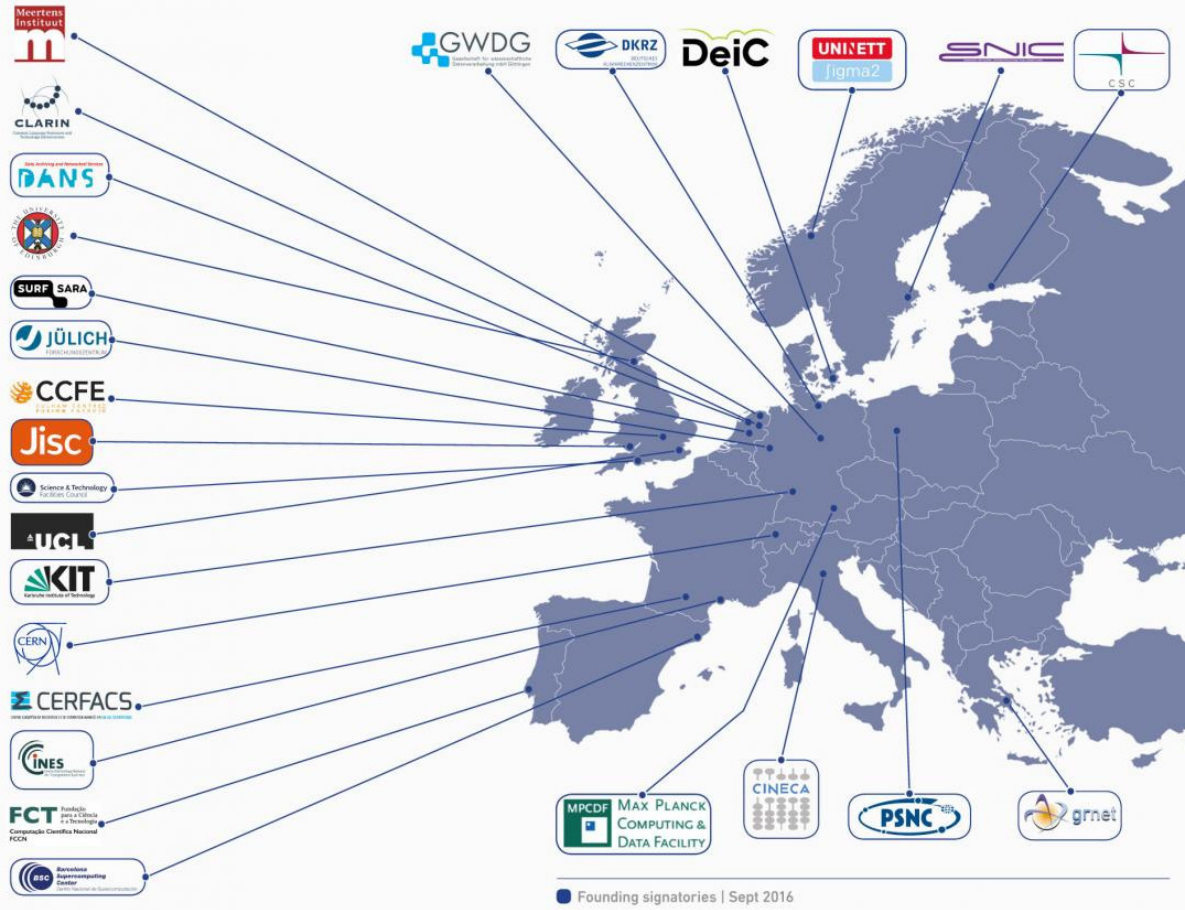
EUDAT(2020) PARTNERS

EUDAT PARTNERS



EUDAT CDI

- Started October 2016
- Agreement among 26 partners
- Major research organisations, data and computing centres
- Agreement: Sustain EUDAT for the next 10 years



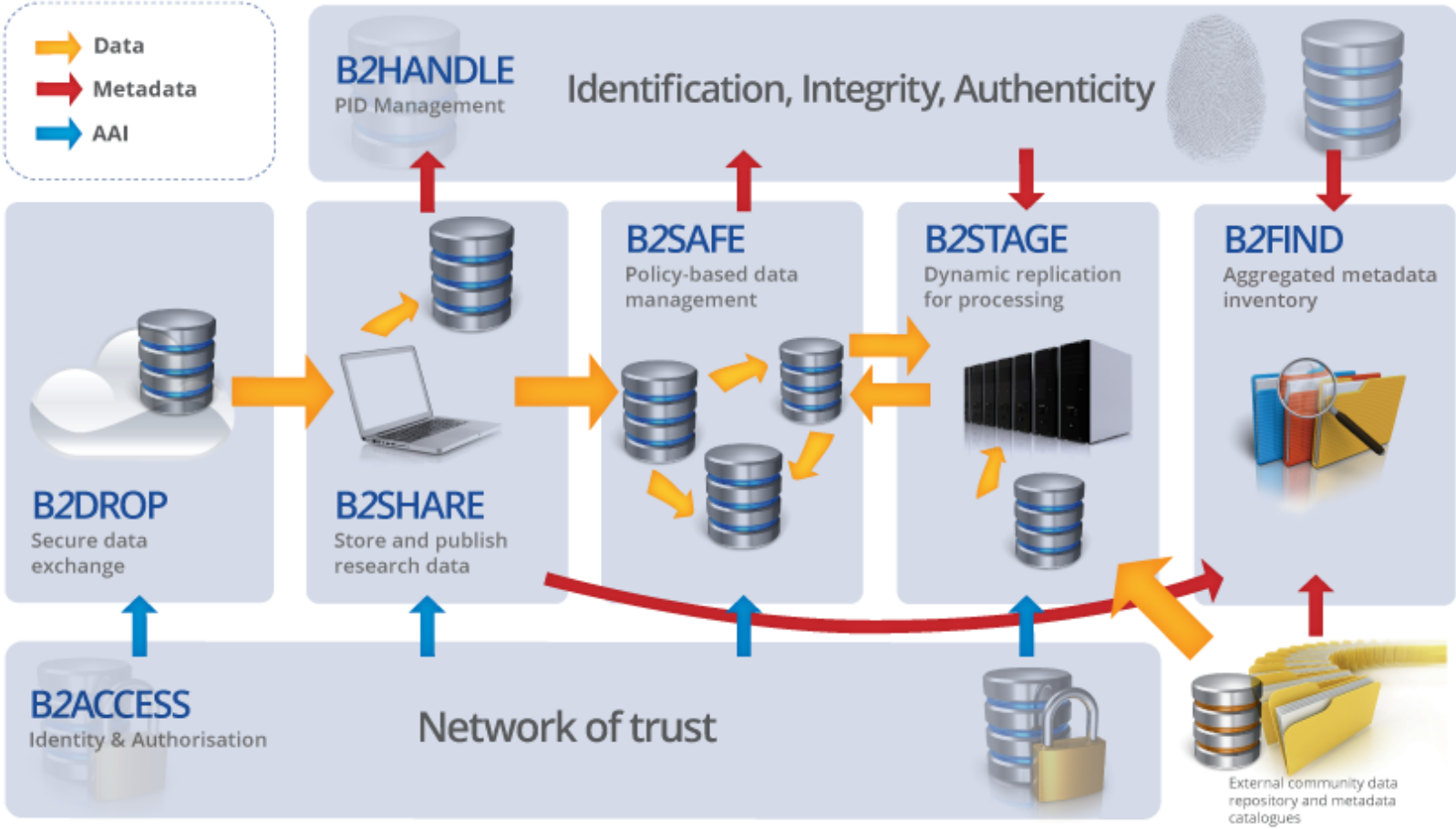
JÜLICH Forschungszentrum

EUDAT IS COMMUNITY DRIVEN

JÜLICH Forschungszentrum

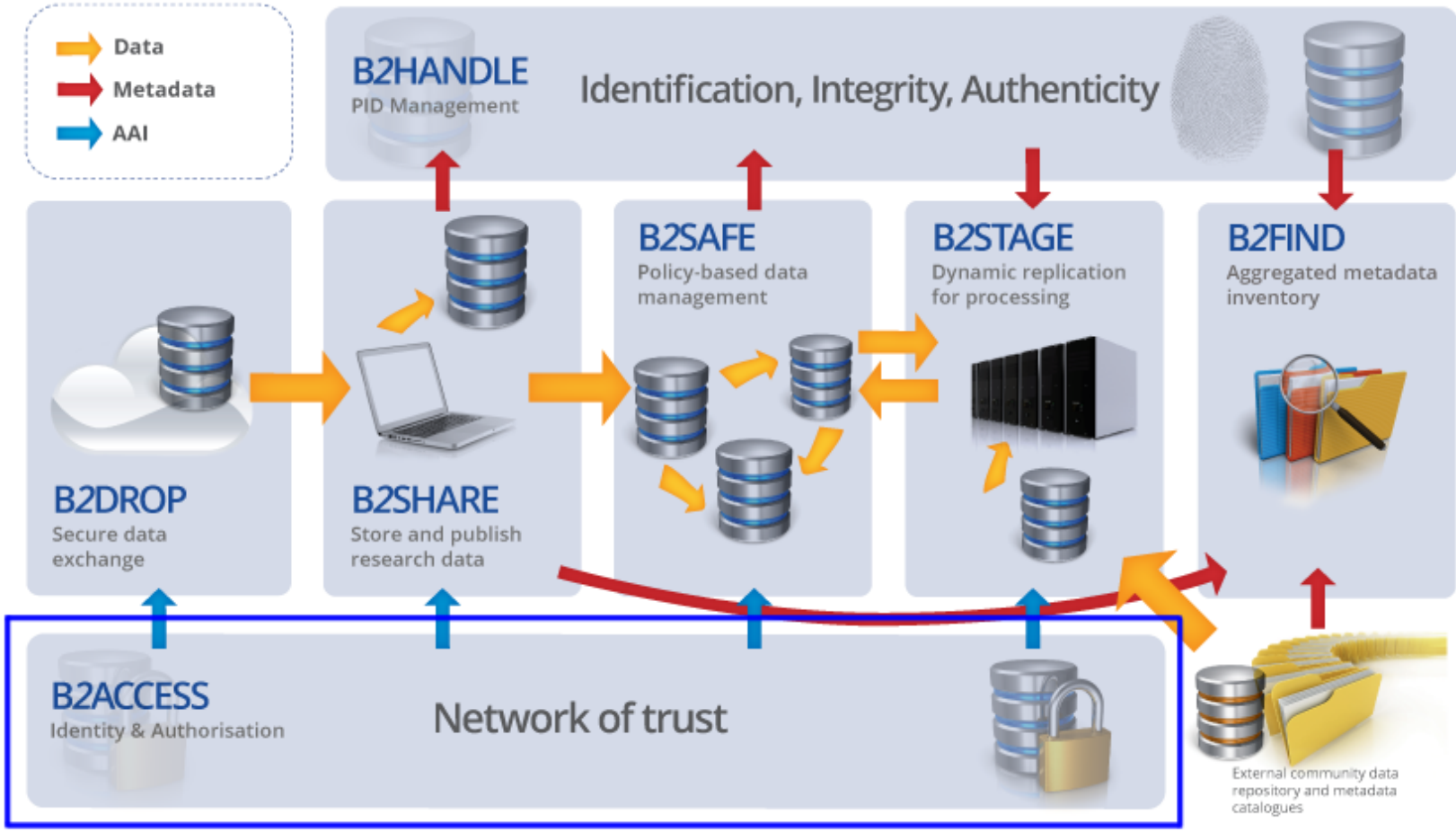
EUDAT: Service Suite

EUDAT SERVICE SUITE - OVERVIEW



JÜLICH Forschungszentrum

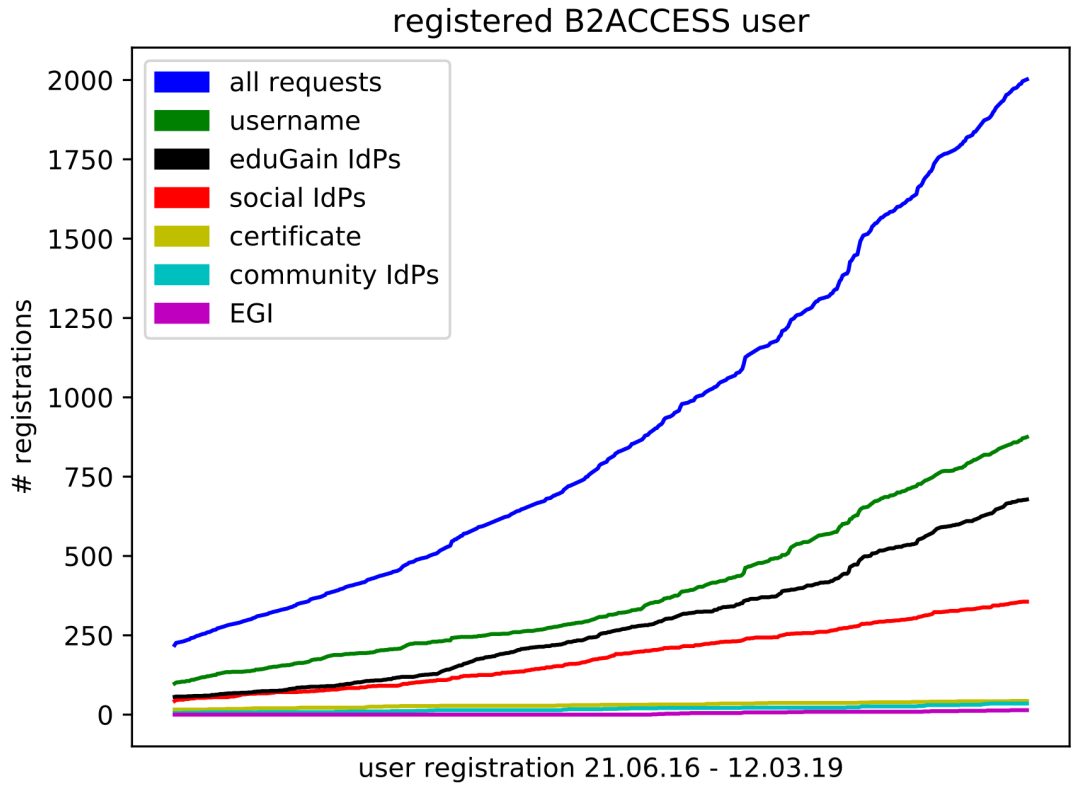
EUDAT SERVICE SUITE - B2ACCESS



JÜLICH Forschungszentrum

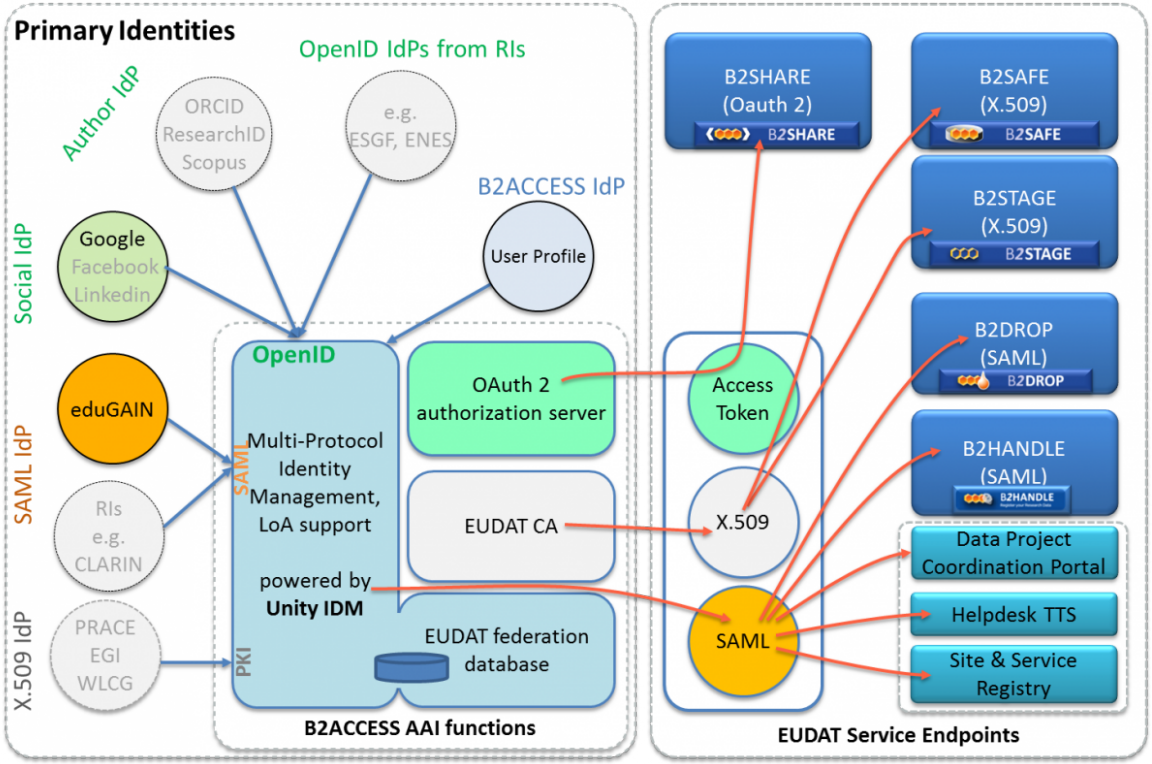
B2ACCESS

- EUDAT Authentication and Authorization Infrastructure
- Based on Unity 2.4.2
- CA for short lived X.509 (testing)
- Acts as a proxy IdP
- Users: 2002



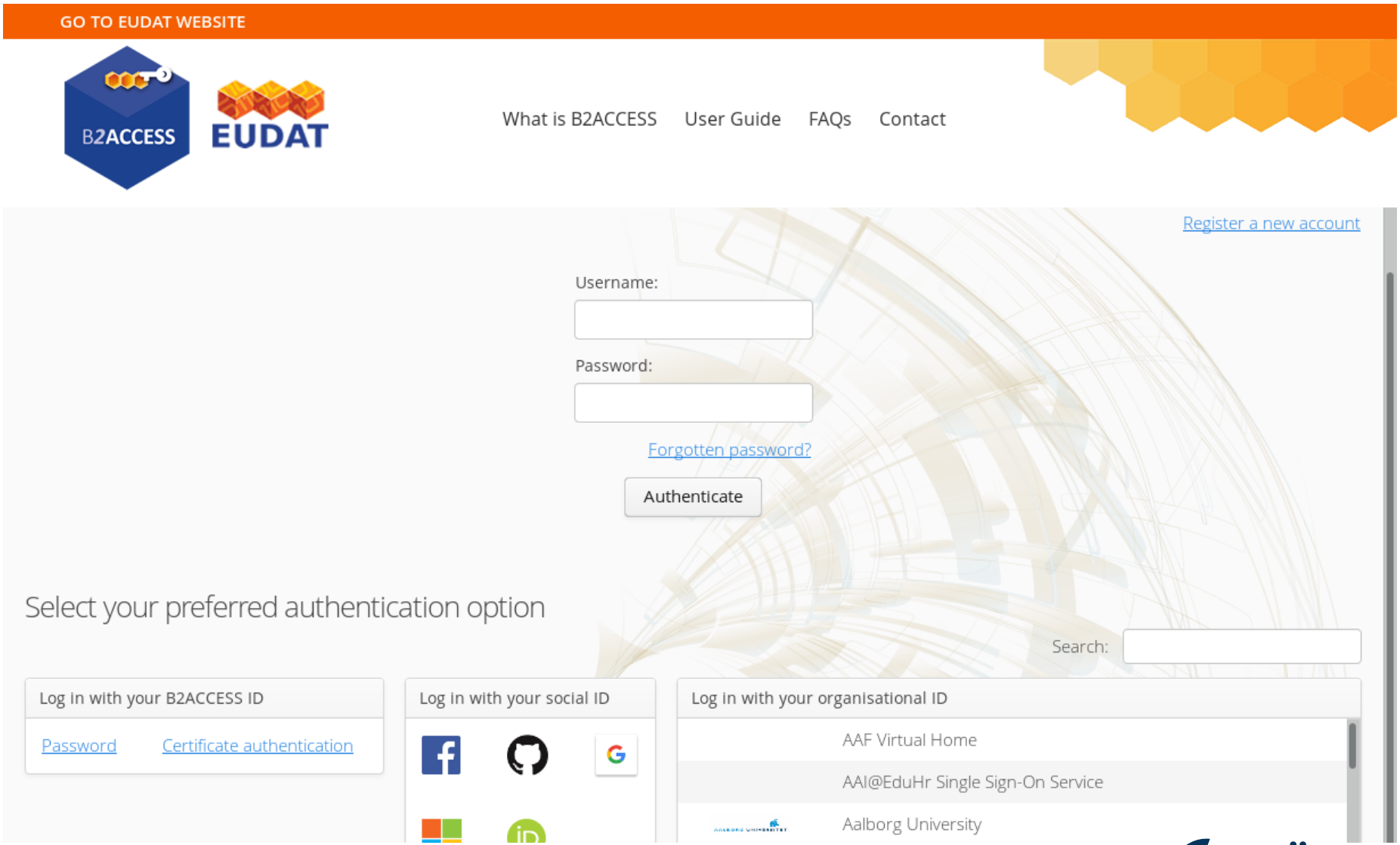
B2ACCESS - ACCOUNTS

- eduGain (SAML)
- Community IdPs (SAML)
  - ARIA
  - Clarin
  - EGI
  - ELIXIR
- Social IdPs (OAuth2)
  - Facebook
  - GitHub
  - Google
  - Microsoft Live
  - ORCID
- Local B2ACCESS accounts
  - Username
  - Certificate with optional username

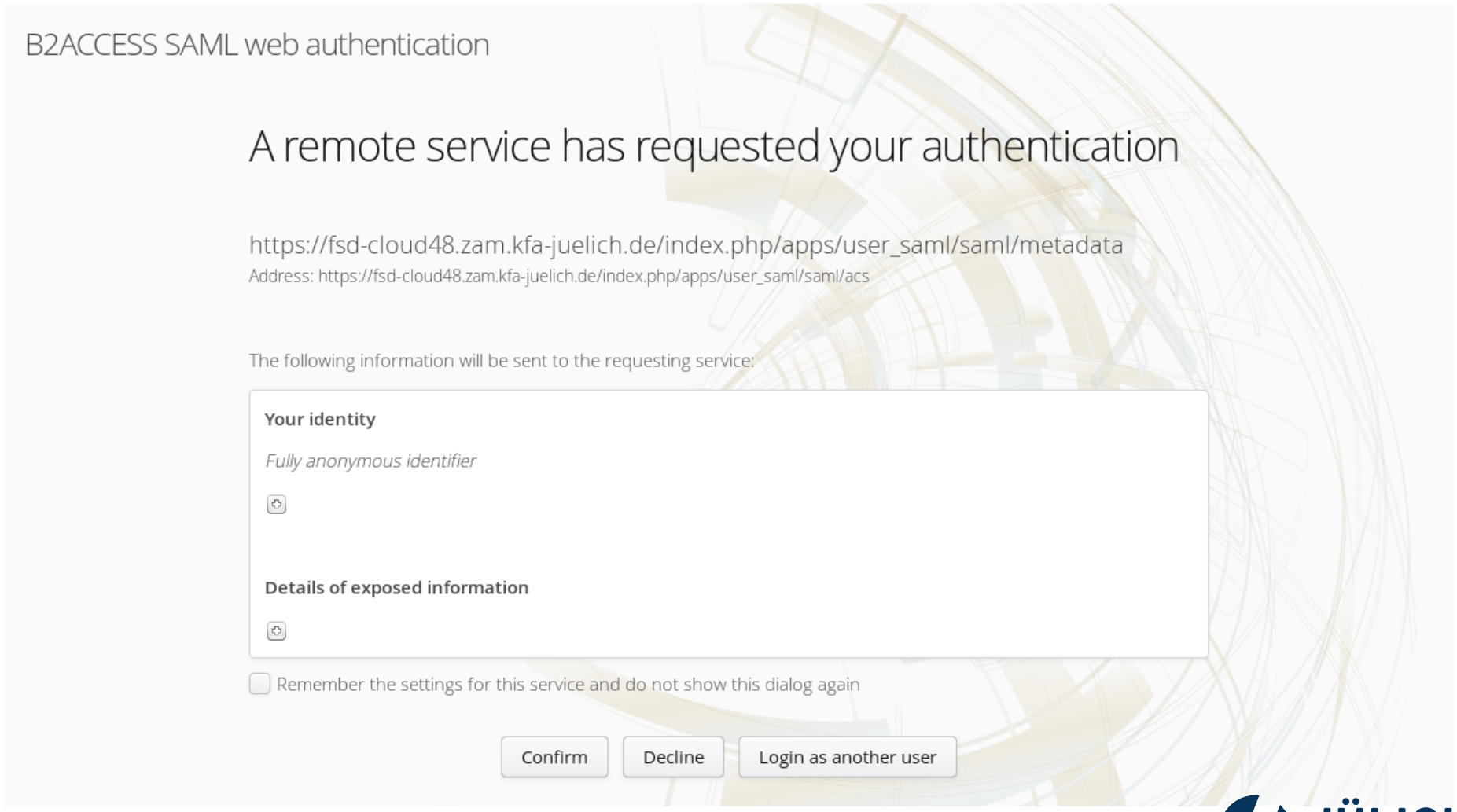


JÜLICH Forschungszentrum

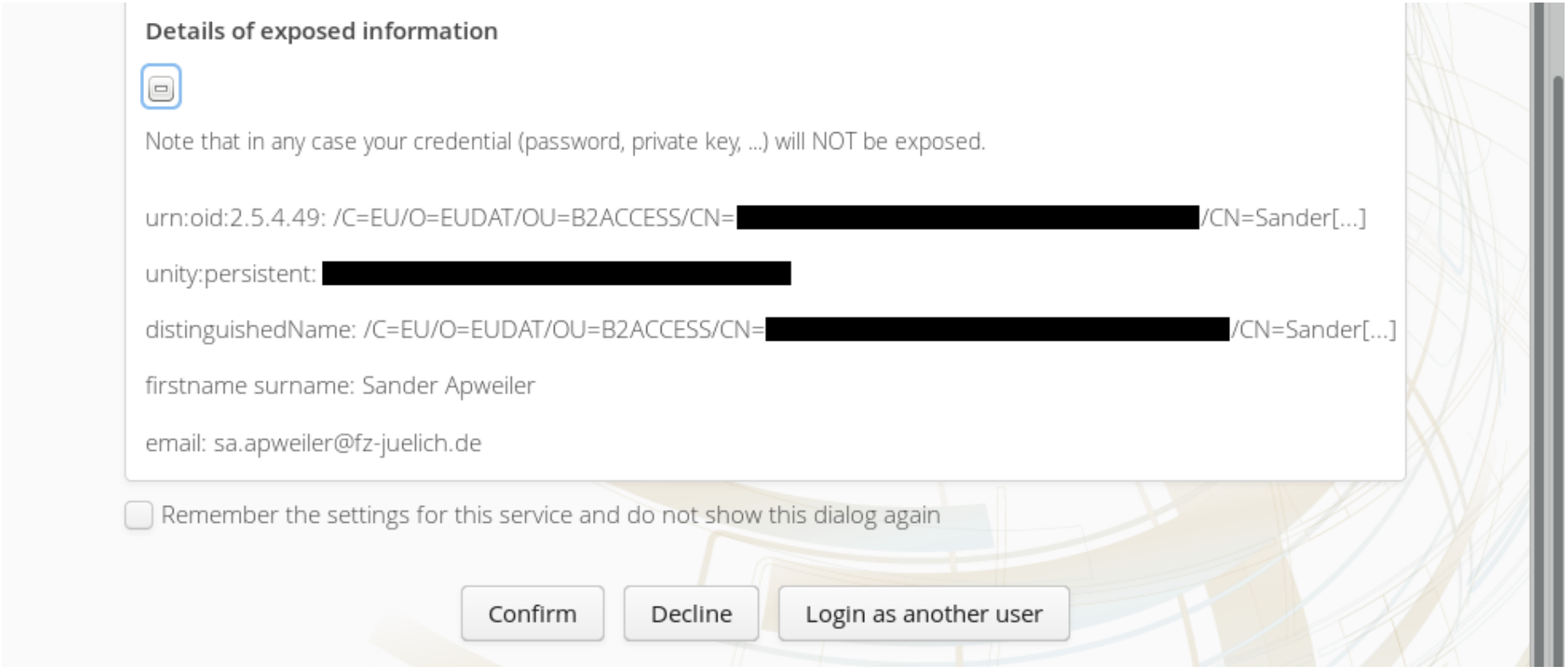
B2ACCESS - LOGIN SCREEN



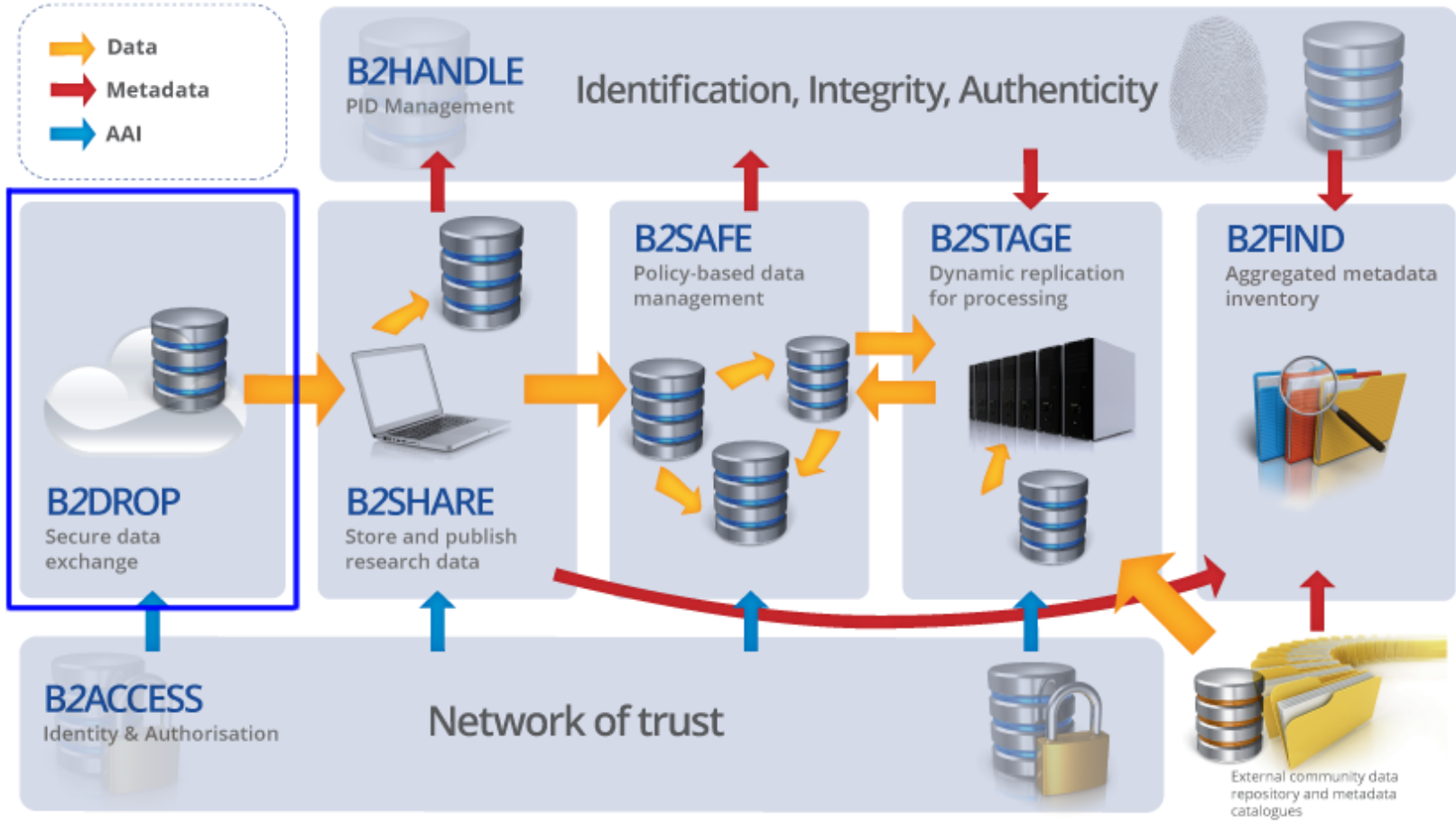
B2ACCESS - ATTRIBUTE RELEASE SCREEN



B2ACCESS - ATTRIBUTE RELEASE SCREEN



EUDAT SERVICE SUITE - B2DROP



JÜLICH Forschungszentrum

B2DROP

For whom?

Researchers (including citizen scientists) who want to share their research with a small set of collaborators prior to making it available to a wider audience.

What can you do with it?

- A service to store and exchange data
- Automated desktop synchronization on many platforms

B2DROP

For whom?

Researchers (including citizen scientists) who want to share their research with a small set of collaborators prior to making it available to a wider audience.

What can you do with it?

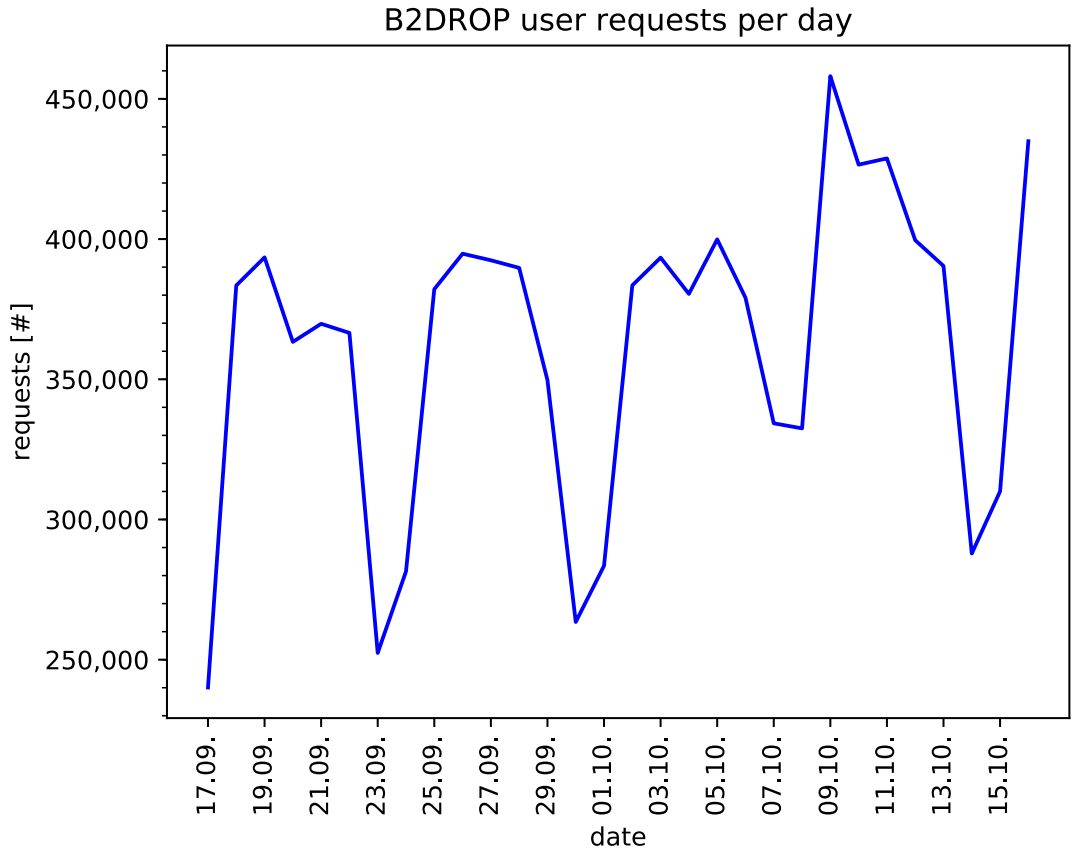
- A service to store and exchange data
- Automated desktop synchronization on many platforms
- Move data to other EUDAT services

B2DROP - SOFTWARE

- Based on Nextcloud
  - Currently 13.0.5
- New theme for EUDAT CI
- New app called B2SHAREBRIDGE
  - Integration with B2SHARE
- user_saml module
  - Integration with B2ACCESS
  - Some own modifications

B2DROP - USAGE STATISTICS

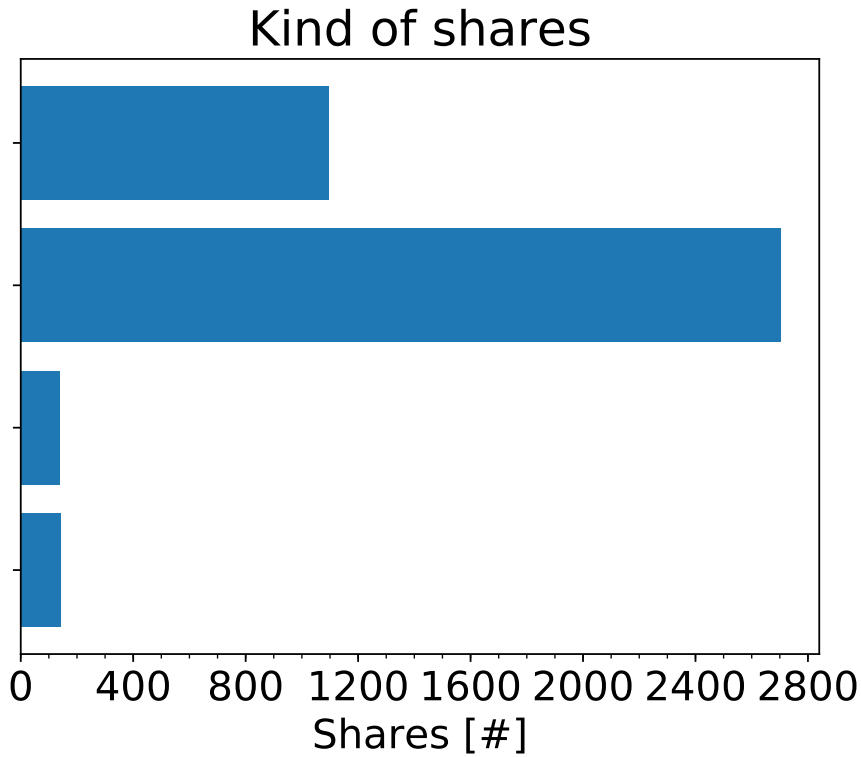
- EUDAT catch all instance
- Users: 1813
  - 1340 in 10/2017
  - 600 in 05/2016
  - 330 in 09/2015
- Volume: 3.5 TB (15 TB available)
- Objects: 2.759.179 #
- Requests:
  - 358.000 #/day
  - 4,1 #/sec
- Daily backup to tape



JÜLICH Forschungszentrum

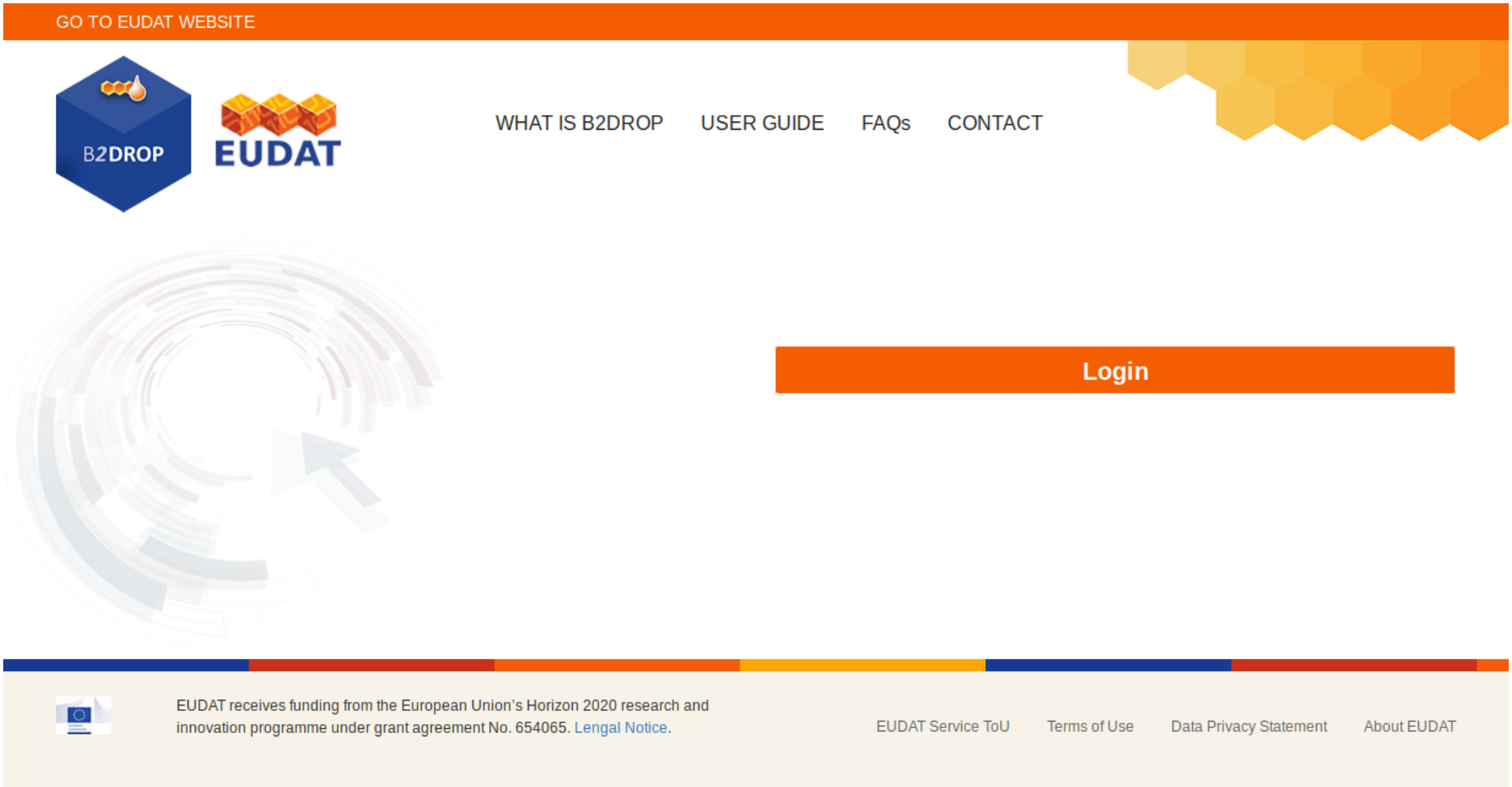
B2DROP - USAGE STATISTICS

- Shares
  - With other B2DROP users
  - By link
  - Federated to other services
  - Federated from other services

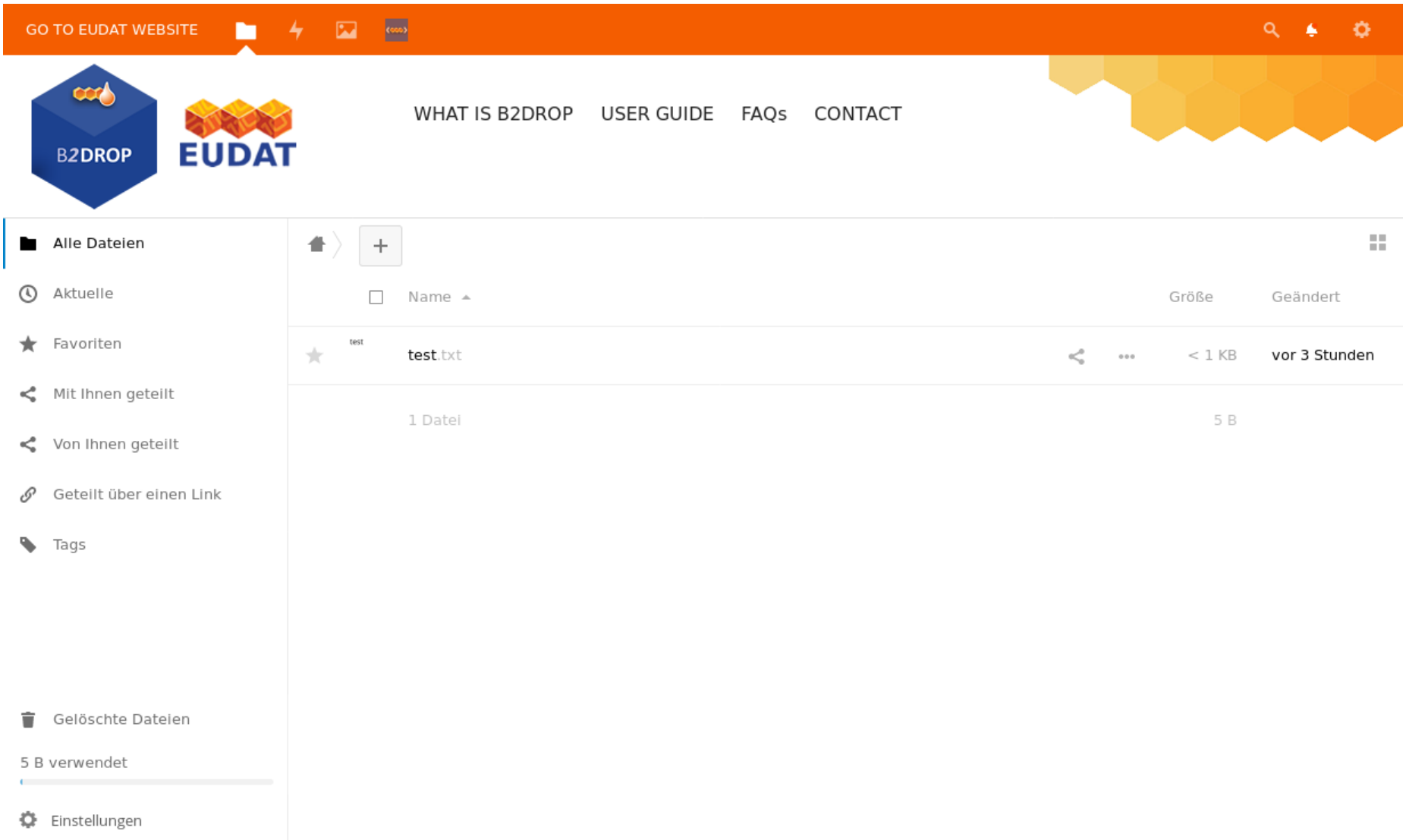


- 71 uploads with B2SHAREBRIDGE

B2DROP - LOGIN SCREEN

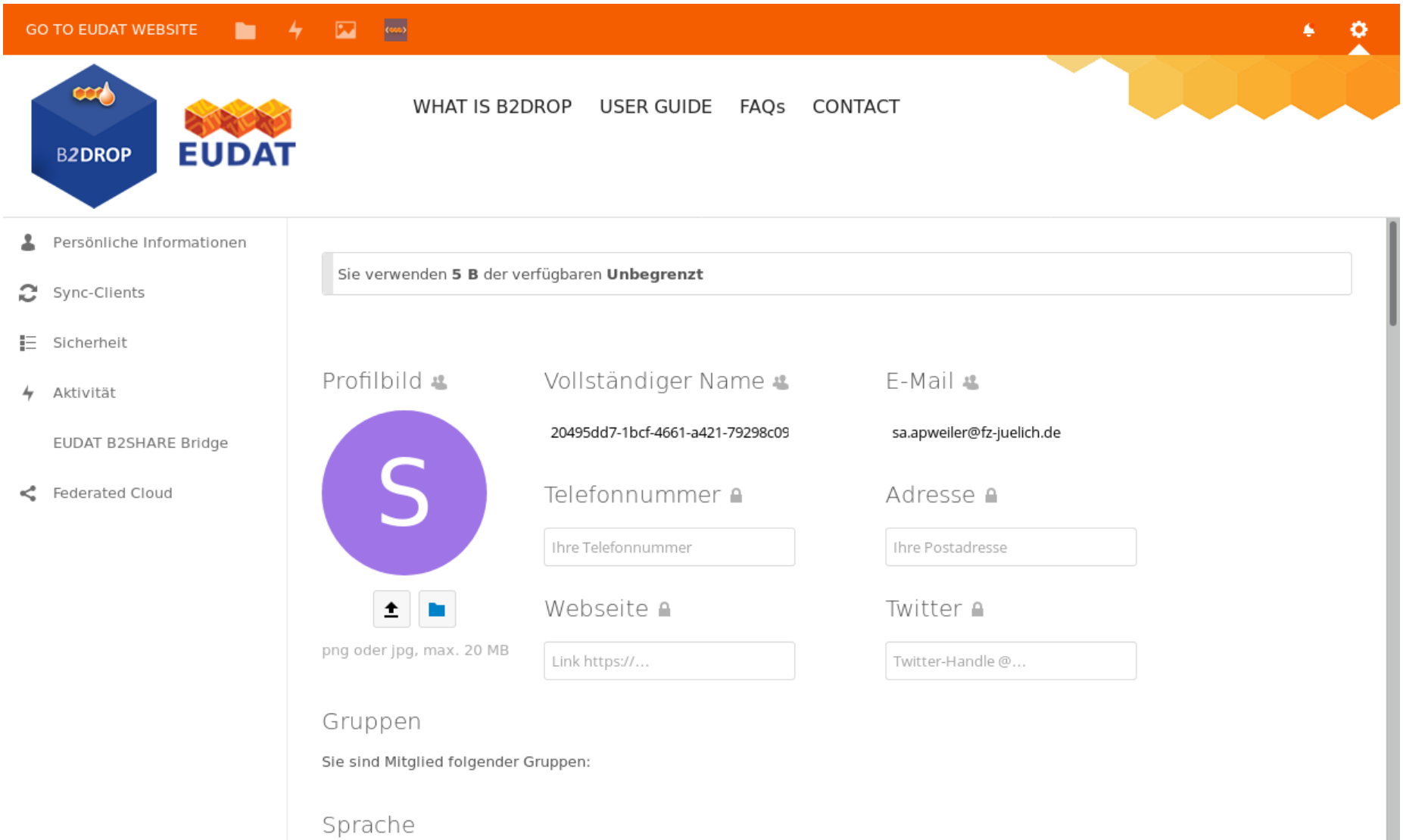


B2DROP - FILE LIST



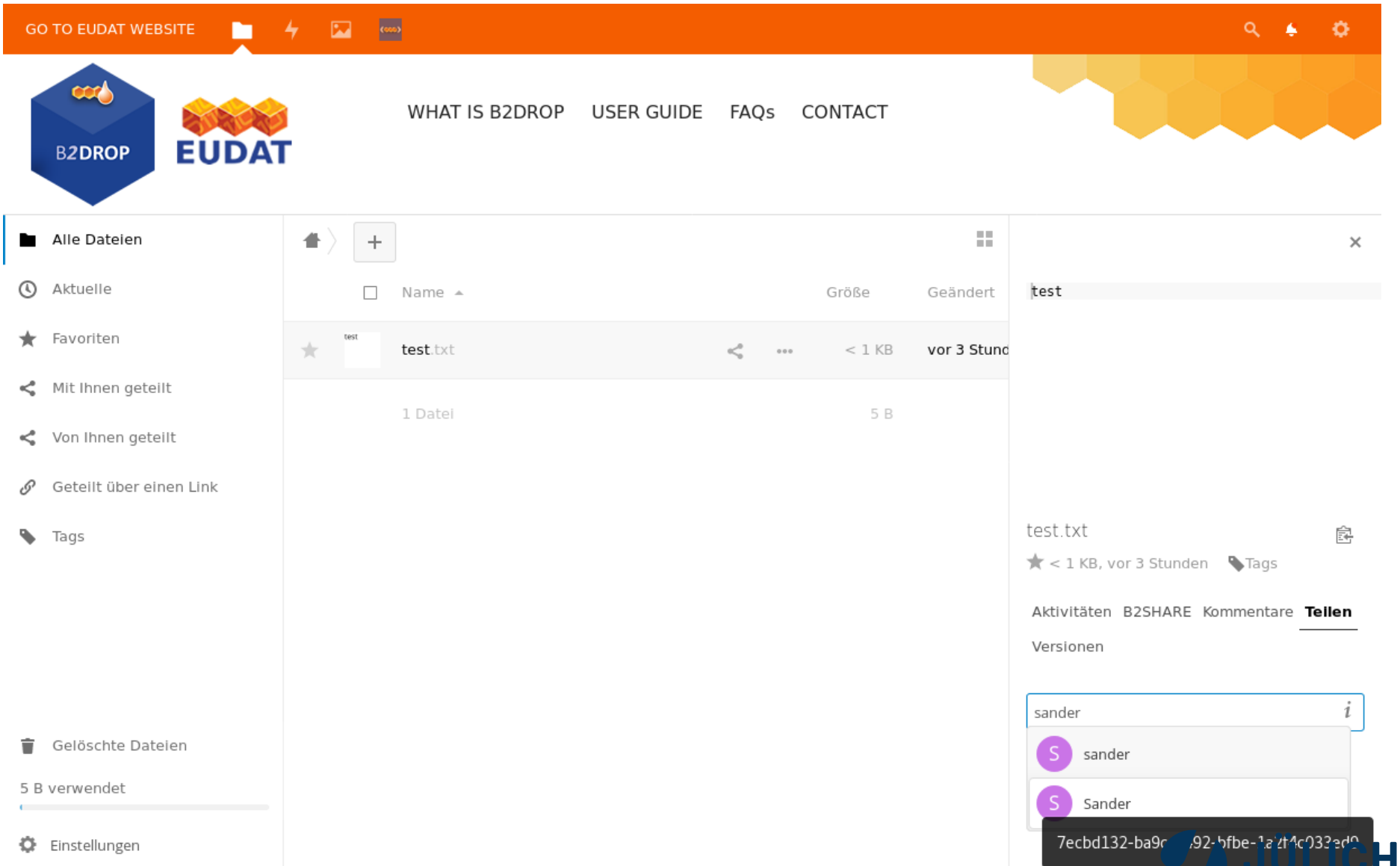
JÜLICH Forschungszentrum

B2DROP - PERSONAL SETTINGS

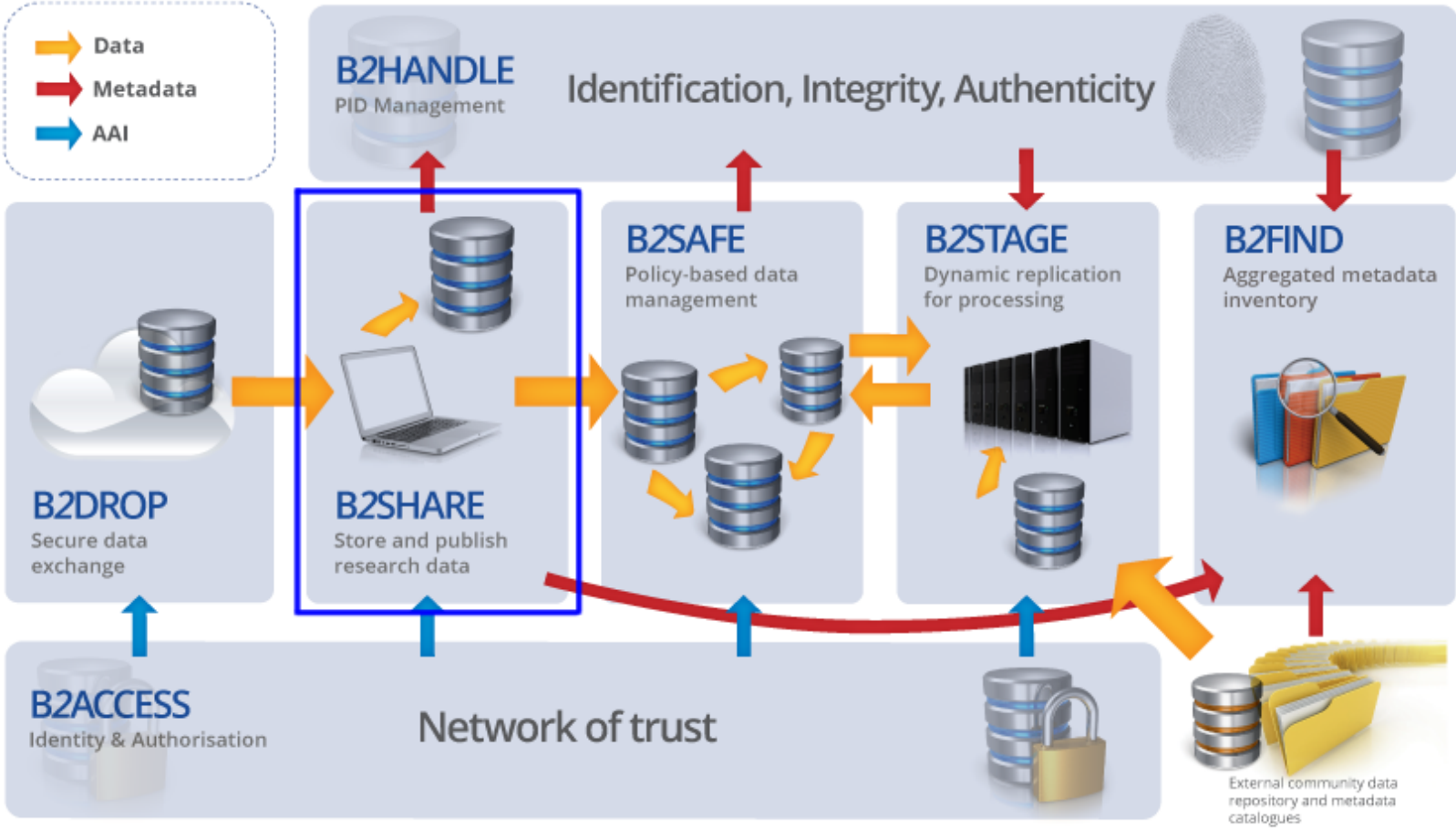


JÜLICH Forschungszentrum

B2DROP - FILE SHARING



EUDAT SERVICE SUITE - B2SHARE



JÜLICH Forschungszentrum

B2SHARE

For whom?

Researchers (including citizen scientists) who want to publish their research data to a wider audience and get a unique identifier as a reference to the data.

What can you do with it?

- A service to publish and exchange data
- Store additional (community specific) metadata
- Reference the data with a PID and/or DOI

B2SHARE

For whom?

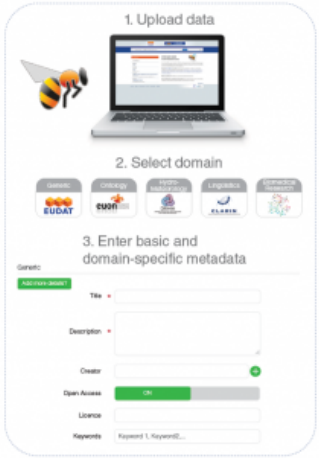
Researchers (including citizen scientists) who want to publish their research data to a wider audience and get a unique identifier as a reference to the data.

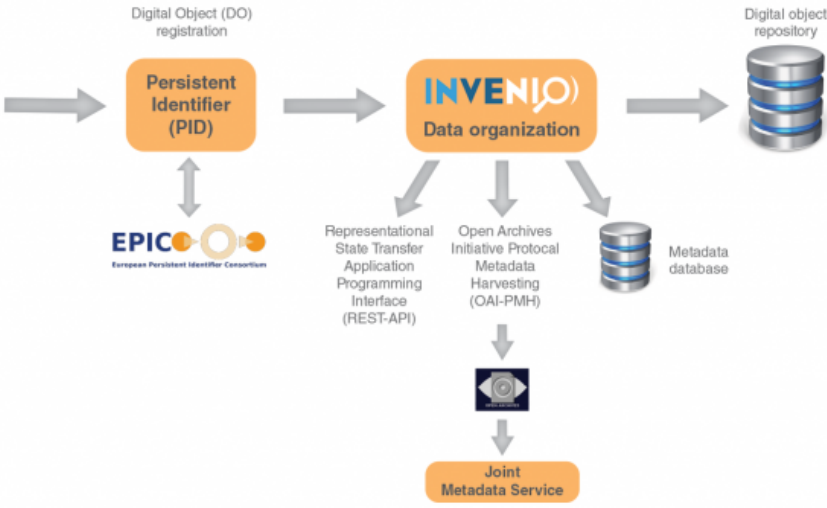
What can you do with it?

- A service to publish and exchange data
- Store additional (community specific) metadata
- Reference the data with a PID and/or DOI
- (Move data to other EUDAT services)

B2SHARE - SOFTWARE

- Based on INVENIO 3
- Own Web UI
- Additional PID/DOI handling
- Deployed with Docker
- Uses OAuth for `B2ACCESS` integration
- Interacts with `B2DROP` and `B2FIND`



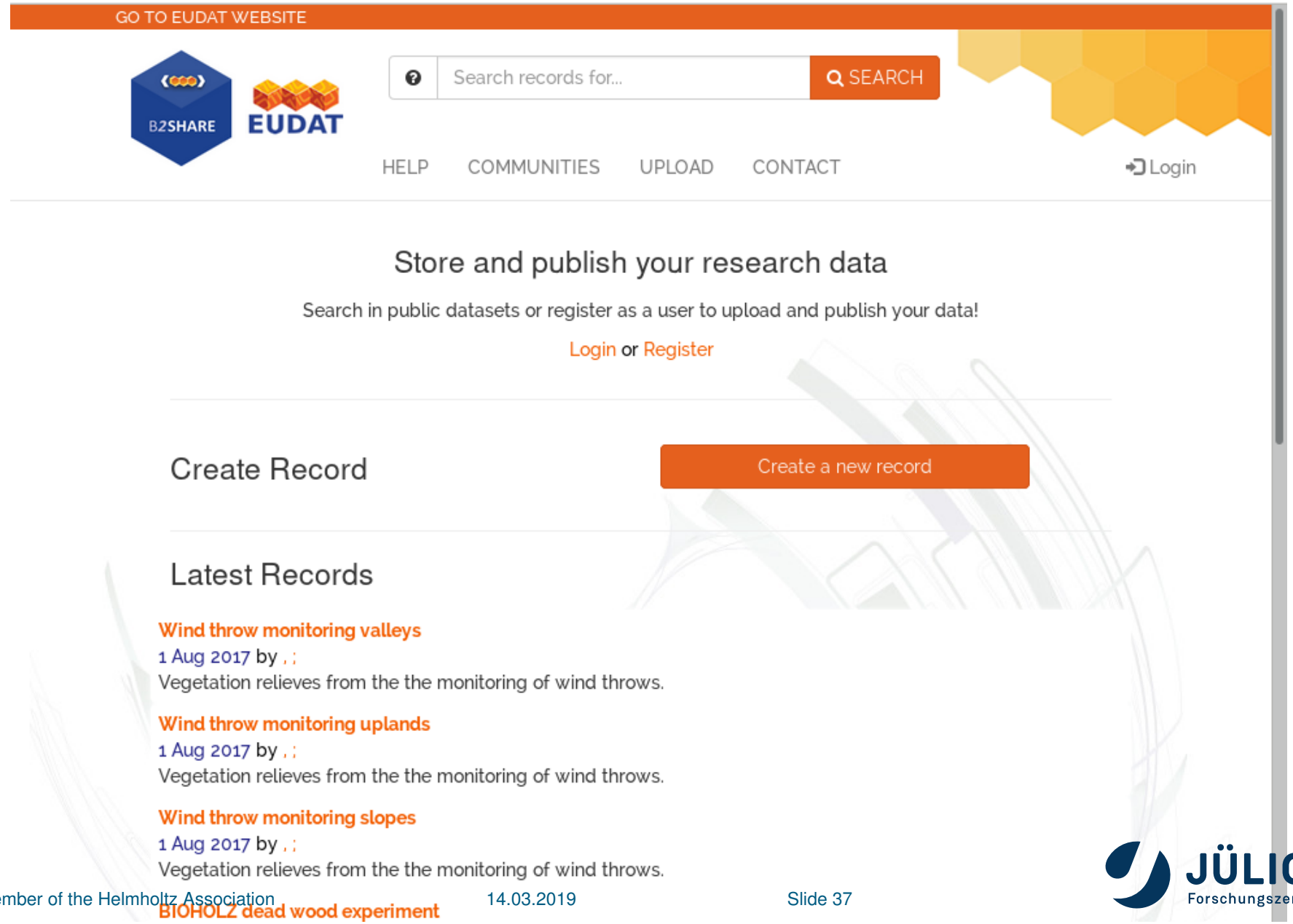


JÜLICH Forschungszentrum

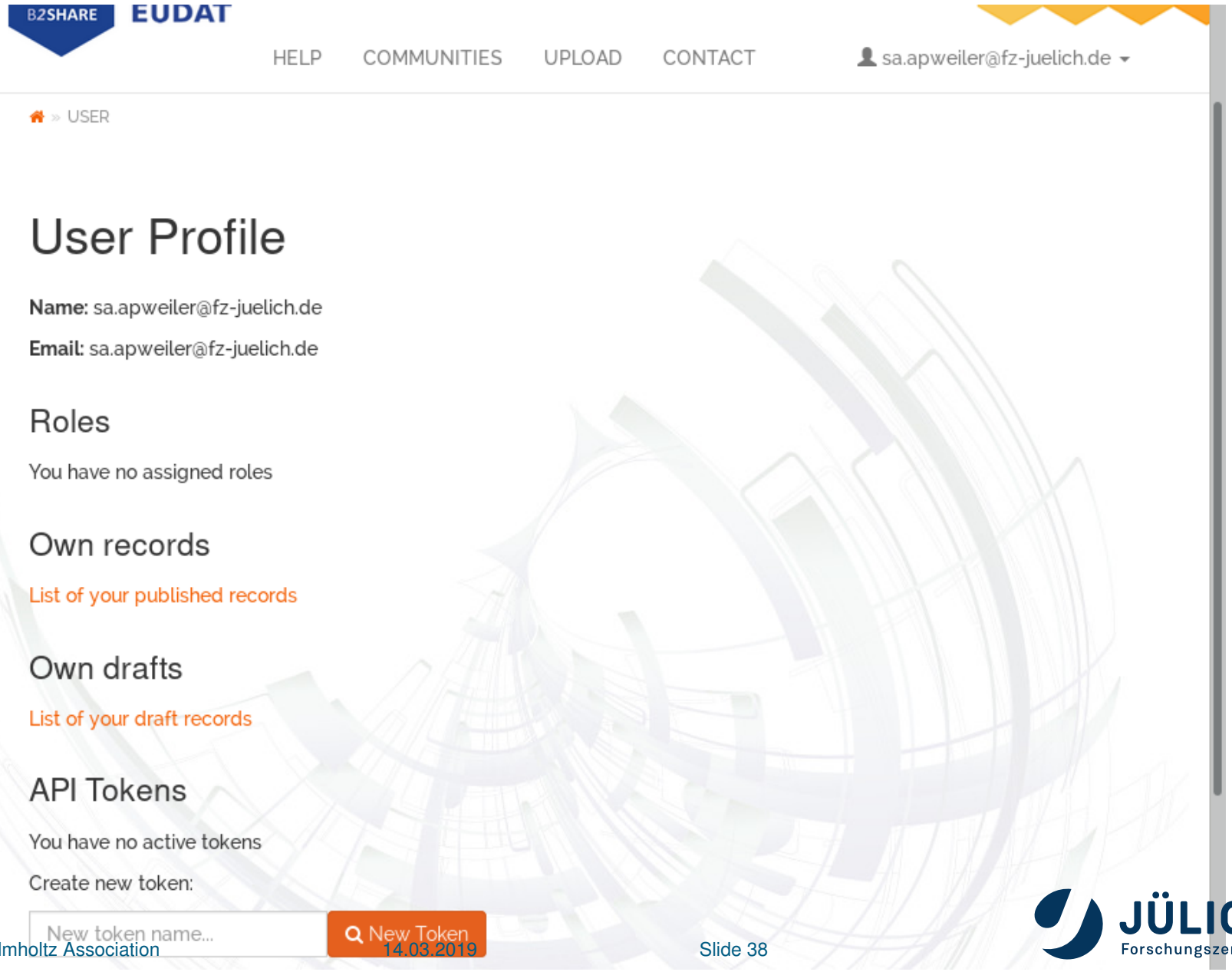
B2SHARE

- Build checksum on files and validates it periodically
  - in case of mismatches the administrator is informed
- File changes by the user create a new version
  - with own PID
  - user can switch between versions
- Records can be deleted by administrators only

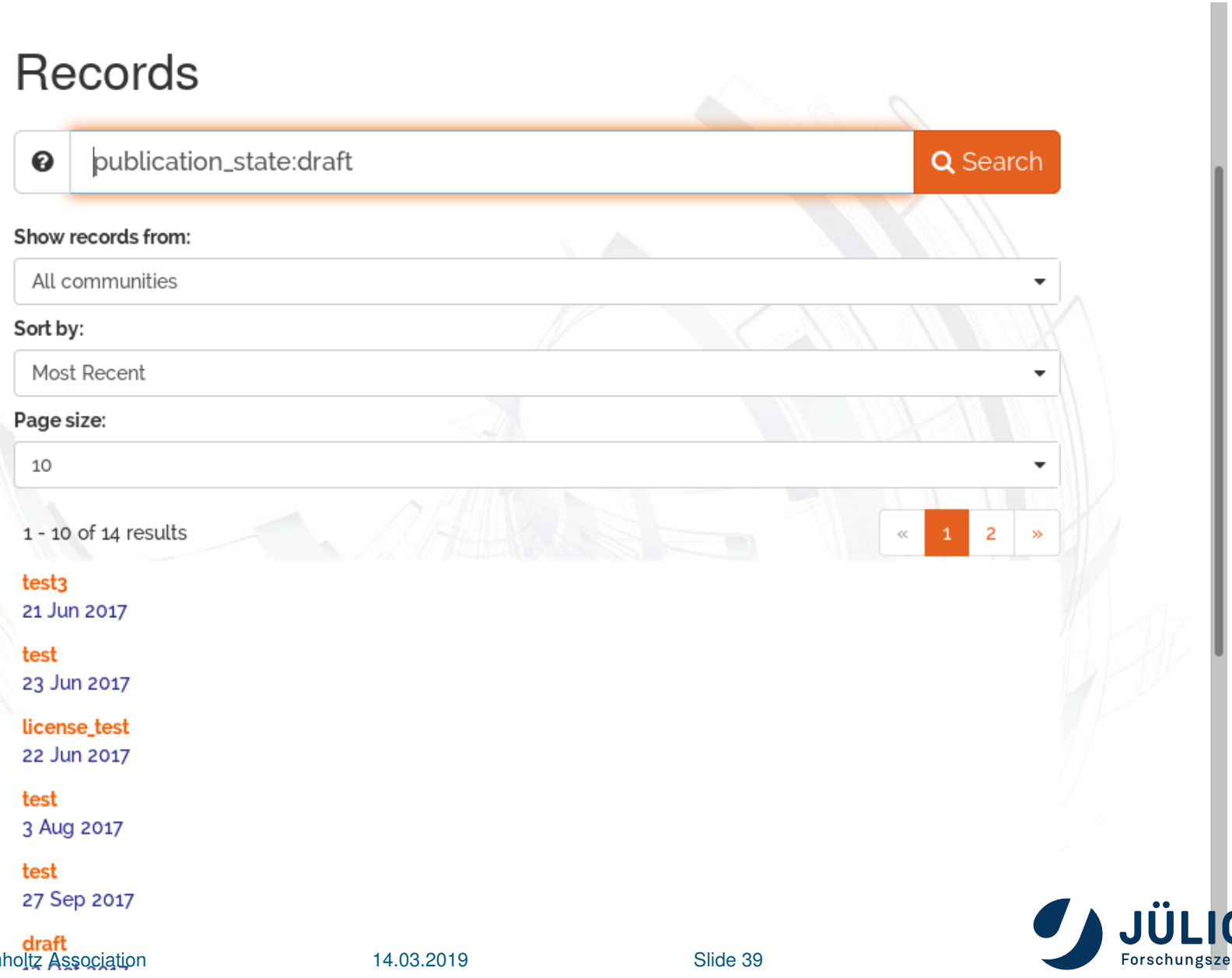
B2SHARE - LOGIN SCREEN



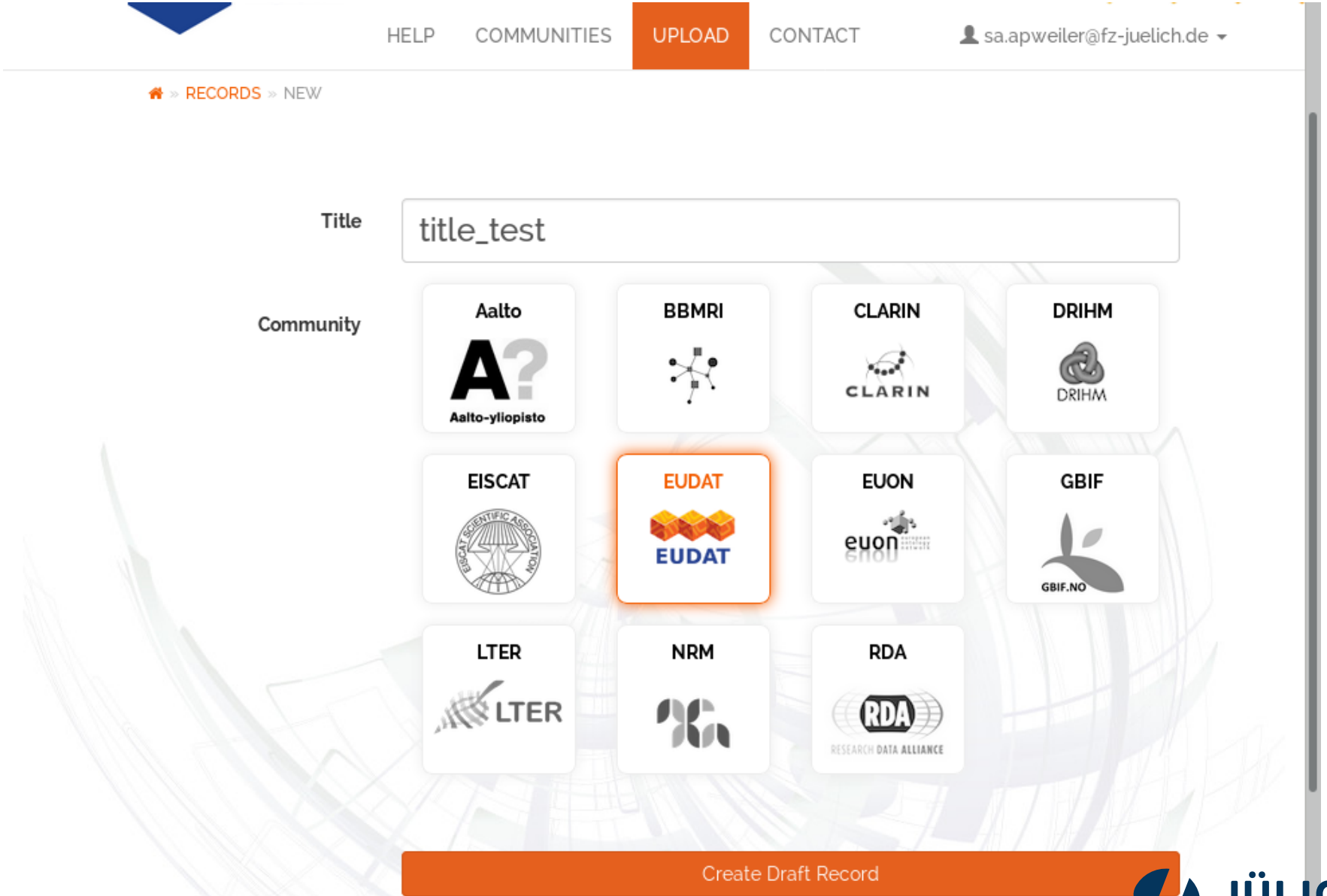
B2SHARE - PERSONAL SETTINGS



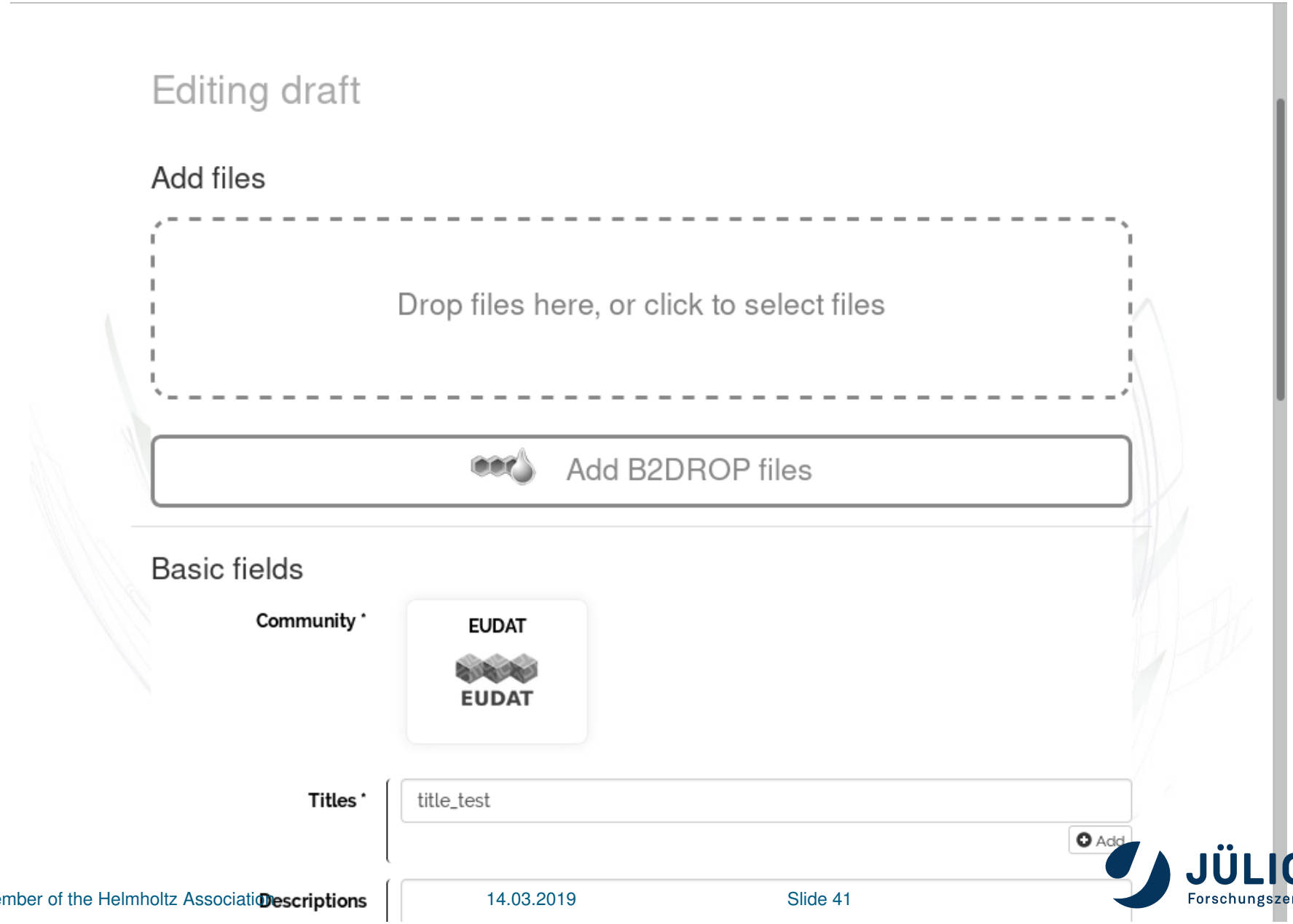
B2SHARE - LIST DRAFTS/SEARCH



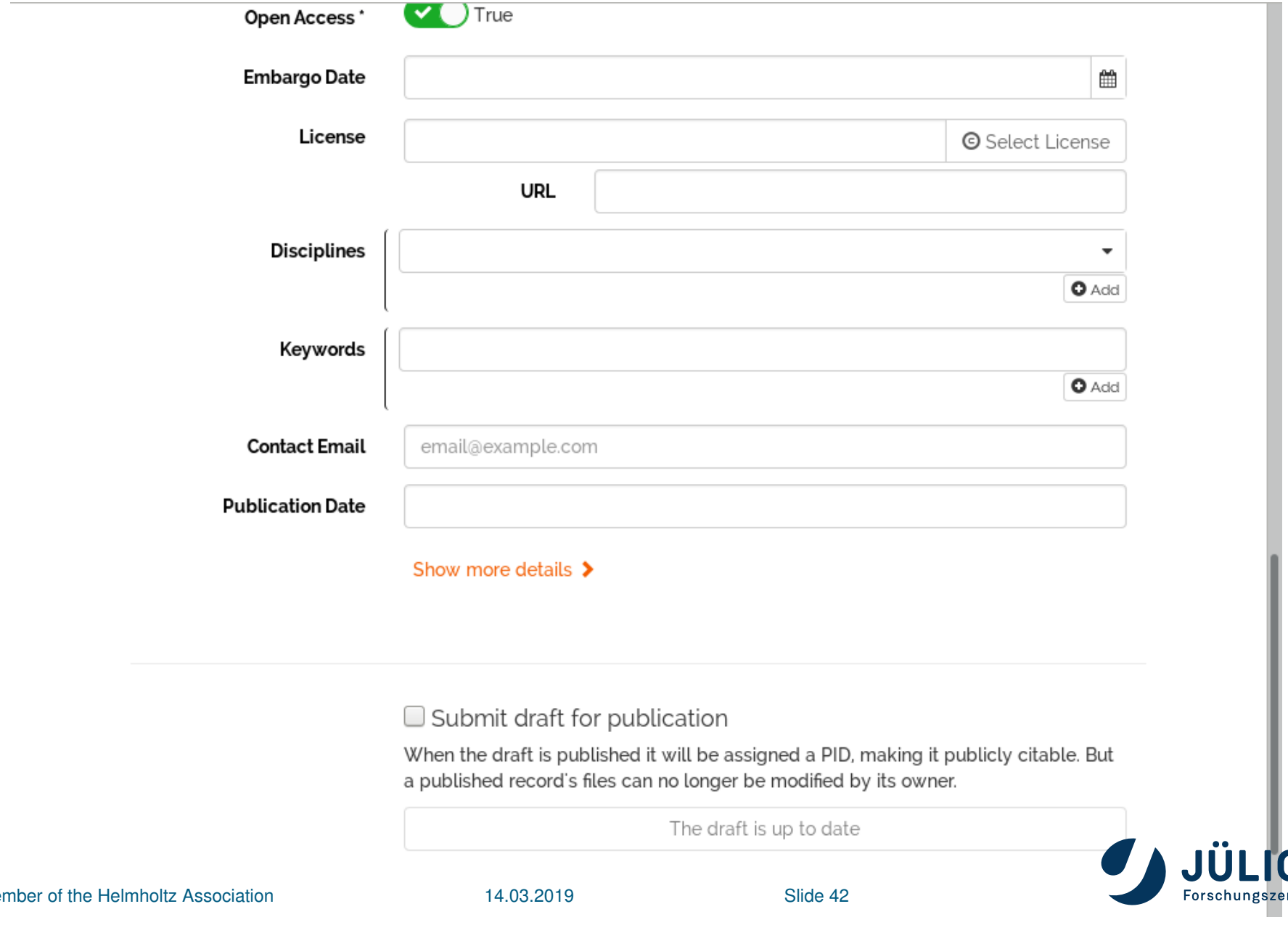
B2SHARE - CREATE A NEW RECORD



B2SHARE - CREATE A NEW RECORD

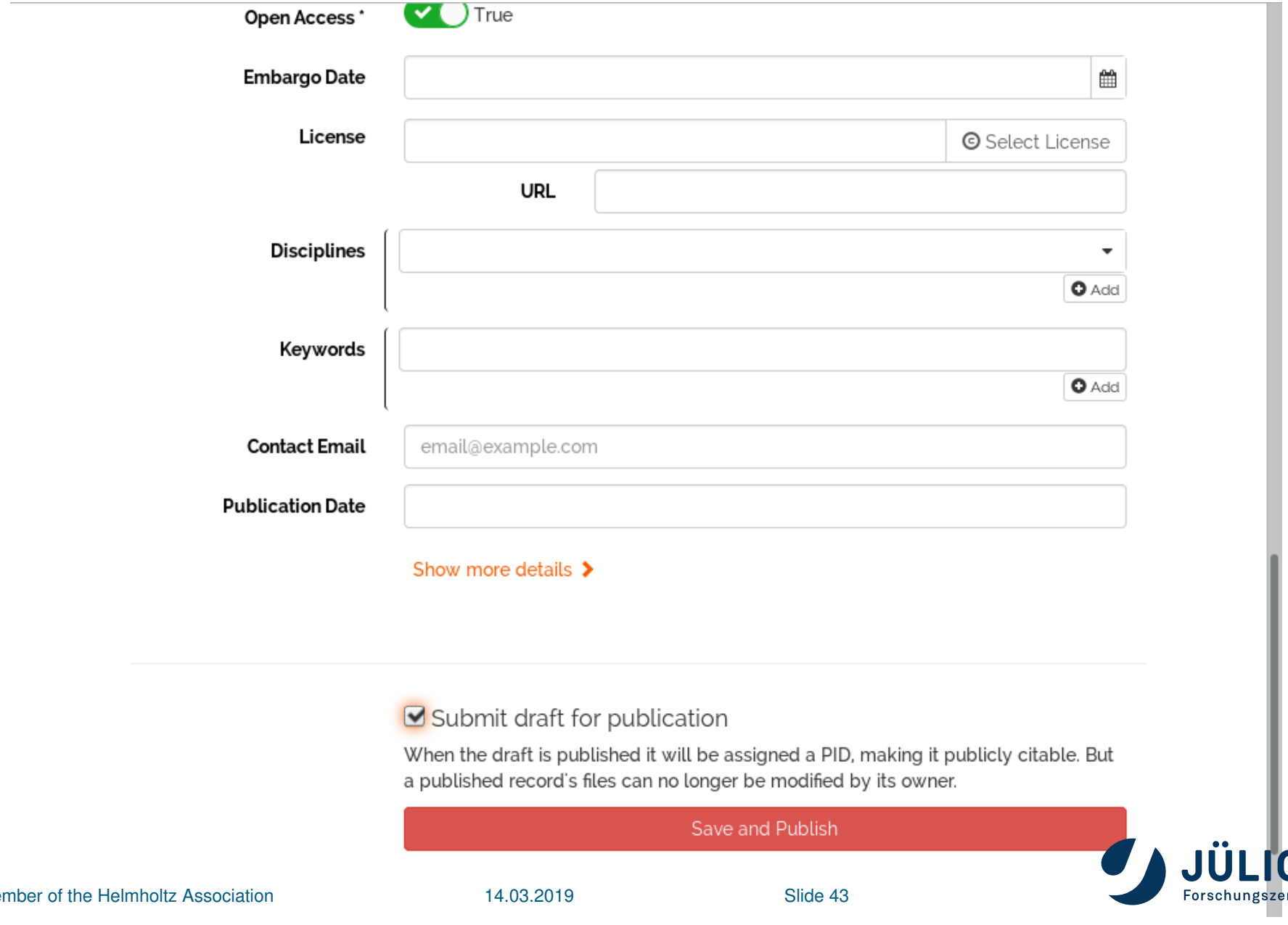


B2SHARE - CREATE A NEW RECORD



JÜLICH Forschungszentrum

B2SHARE - CREATE A NEW RECORD



JÜLICH Forschungszentrum

B2SHARE - API

Upload with web frontend has some cons:

- Limitation in file size
- Inefficient if you create multiple uploads
- Can not be automated

REST-API solve this issue:

- Use favoured tool, e.g. curl
- Can be automated in scripts
- Is not limited to browser timeouts

B2SHARE - API COMMANDS

Commands for

- Listing all
  - communities
  - records
  - records of a community
  - files from a record
- Fetching
  - a community schema
  - the metadata of a specific record
- Searching for
  - records
  - your drafts

B2SHARE - API COMMANDS

Commands for

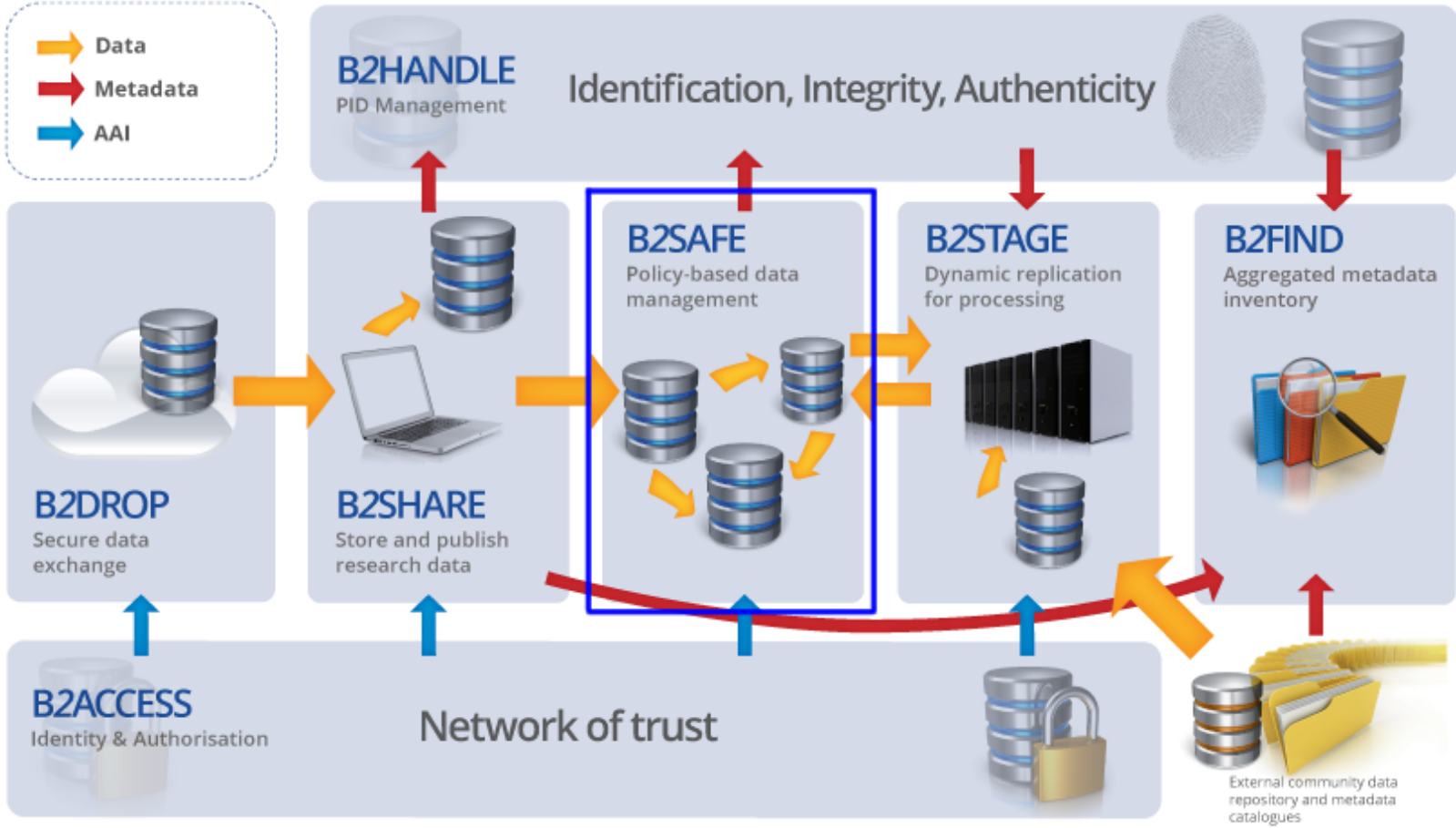
- Uploading
  - a file into your draft
  - metadata into your draft
- Creating a new record
- Deleting a file from your draft
- Submitting a draft for publication
- Reporting abuse
- Sending a request for access to restricted data

B2SHARE - API WORKFLOW

Common API workflow

- Identify a targeted community by listing the available ones
- Use the community identifier to fetch the community metadata schema
- Create a draft
- Upload files to your draft; one command per file
- Set the metadata to your record
- Publish your draft

EUDAT SERVICE SUITE - B2SAFE



JÜLICH Forschungszentrum

B2SAFE

For whom?

Researchers or communities whose repositories lack the capacity and / or funding to offer reliable storage and access services over longer periods of time or without adequate compute capacity.

What can you do with it?

- 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

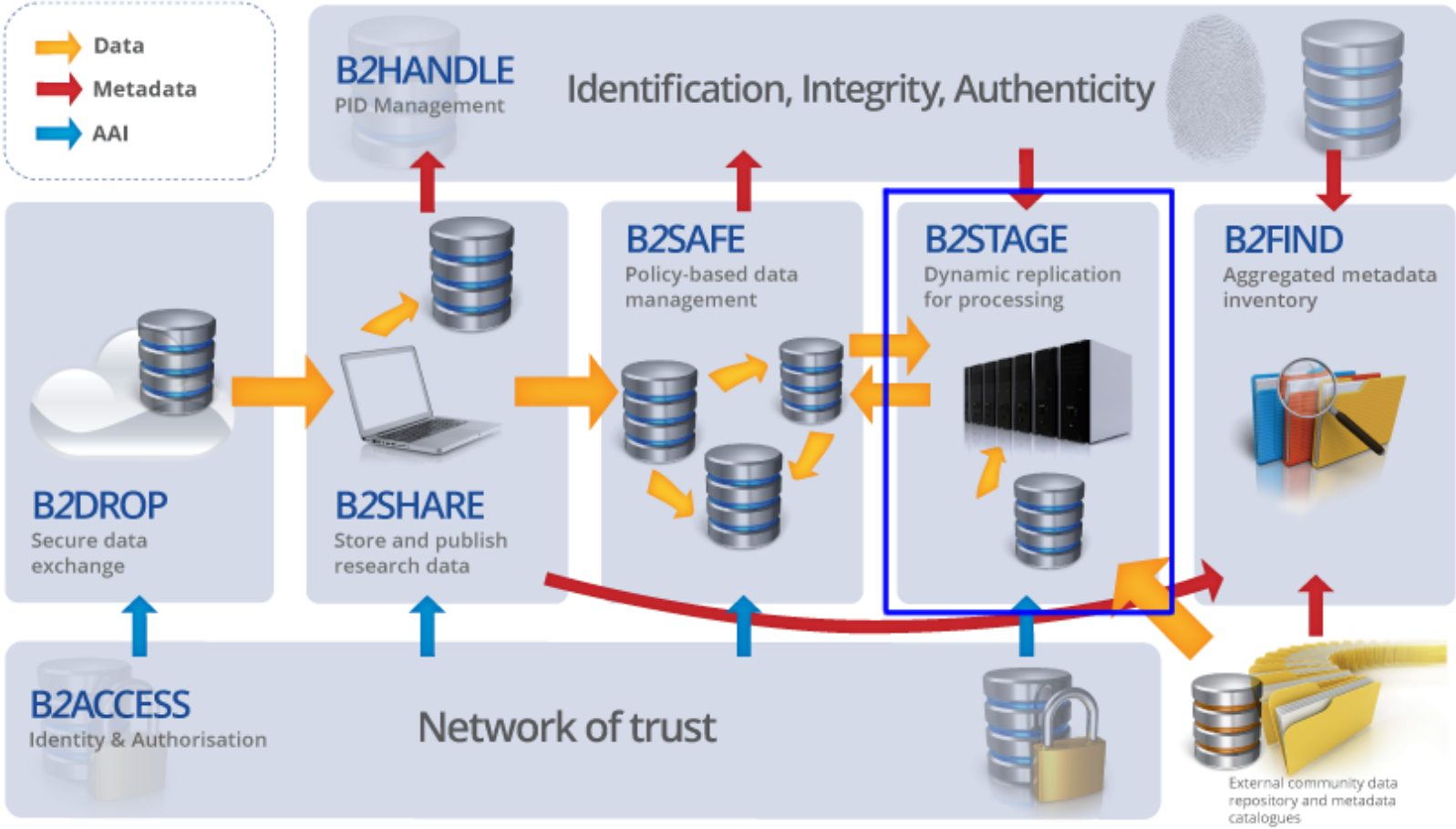
B2SAFE

- Based iRODS
- Own rules for safe replication and PID management
- Use EPIC Handle service

- Each EPIC record knows ”parents” and ”children”

- Starts with replication from community to first EUDAT partner
- Additional replications between EUDAT partners are possible

EUDAT SERVICE SUITE - B2STAGE



JÜLICH Forschungszentrum

B2STAGE

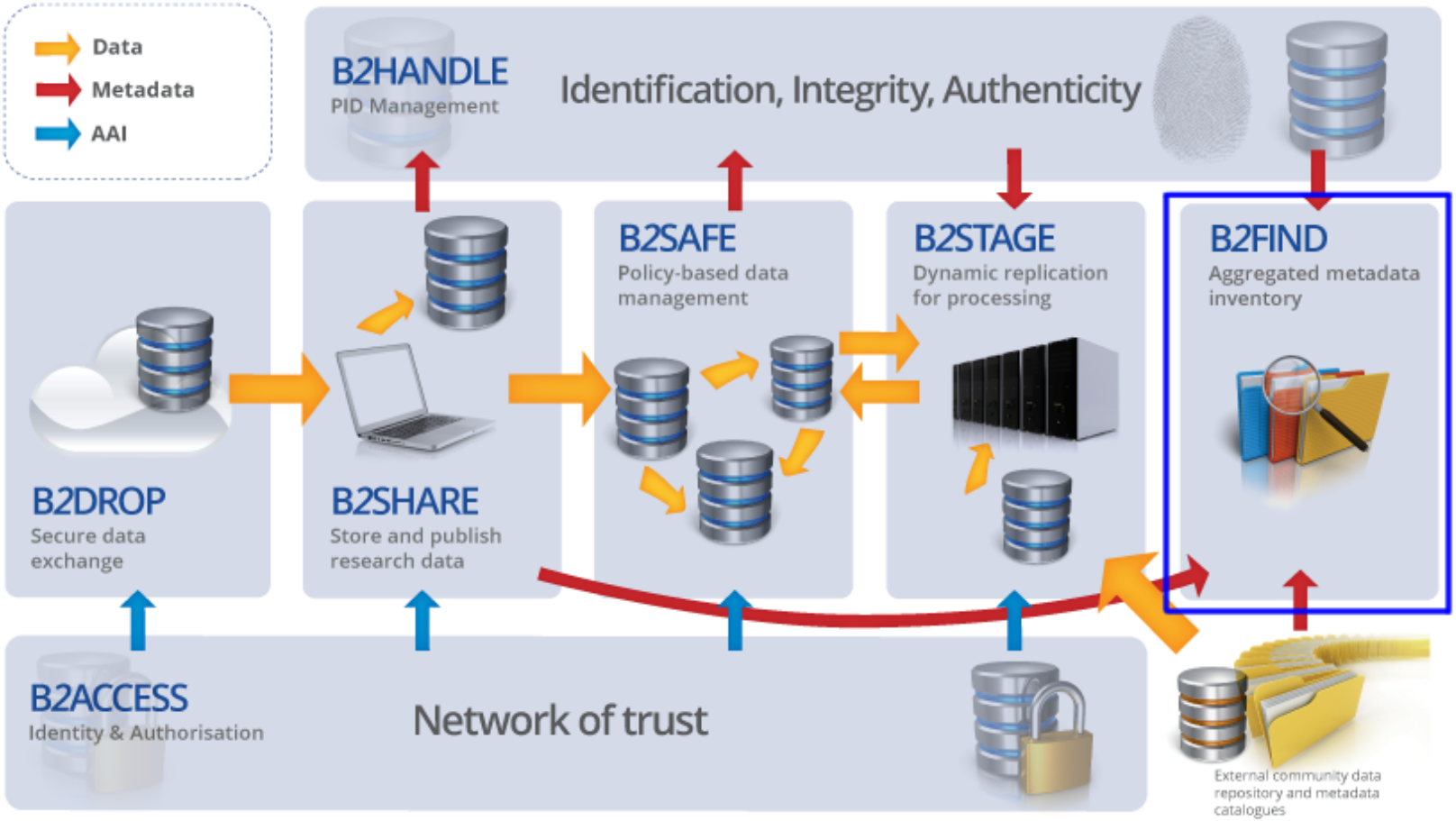
For whom?

Researchers or communities who need to access both large-scale data storage and high-performance computing systems and transfer their data easily between the EUDAT storage resources and remote HPC facilities.

What can you do with it?

- 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

EUDAT SERVICE SUITE - B2FIND



JÜLICH Forschungszentrum

B2FIND

For whom?

Researchers (including citizen scientists) who are searching for specific research data and publications.

B2FIND

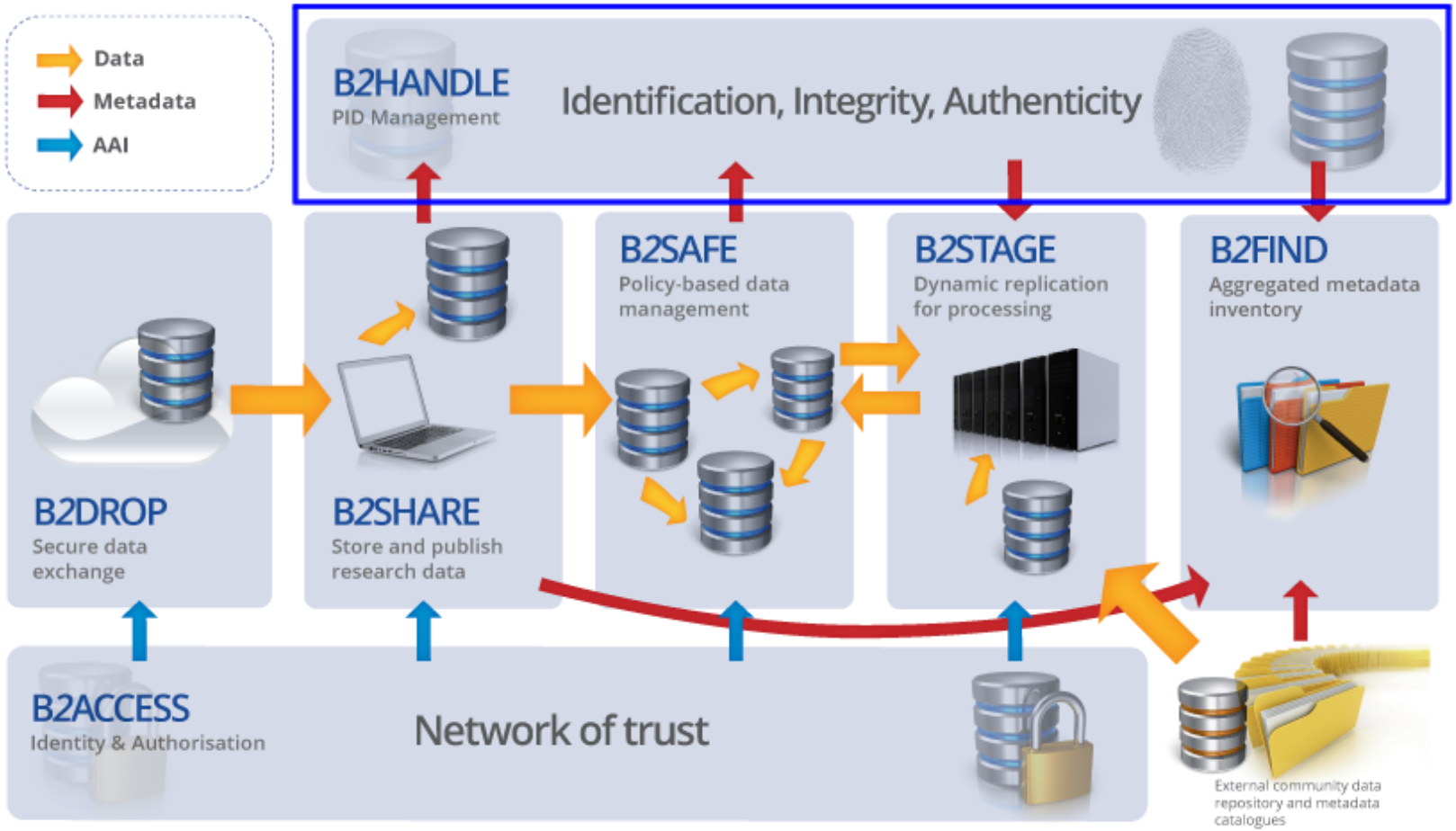
For whom?

Researchers (including citizen scientists) who are searching for specific research data and publications.
Research communities who want to provide their metadata to a wider audience.

What can you do with it?

- 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

EUDAT SERVICE SUITE - B2HANDLE



JÜLICH Forschungszentrum

B2HANDLE

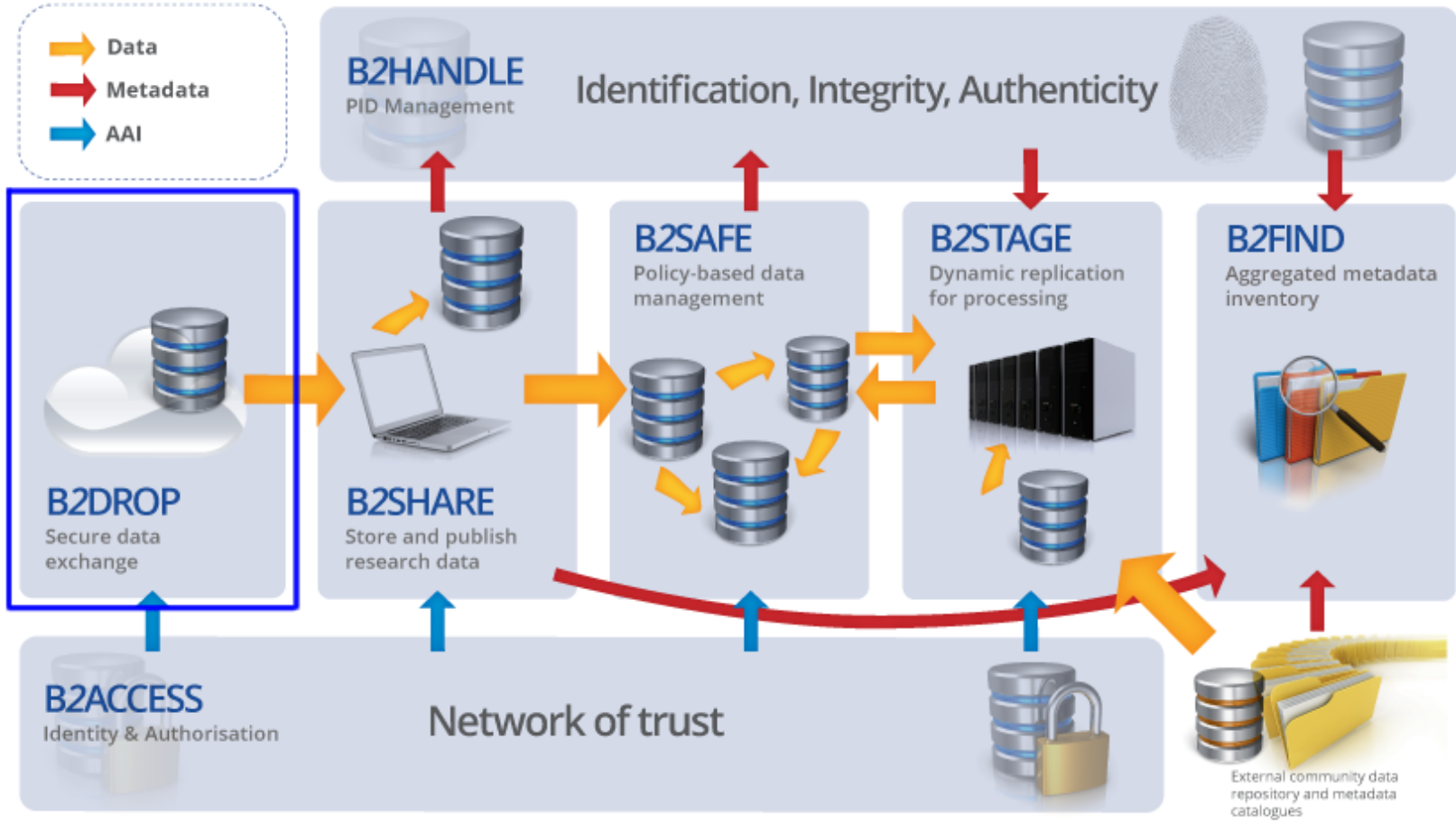
For whom?

Research communities with own data services or data models that are not compliant with the data models in EUDAT services and wish to add PIDs to (data) objects.

What can you do with it?

- Assign PIDs to various kinds of managed objects
- Shielding end-users from the complexity of infrastructure details
- Retrieve and refert to PIDs to objects

EUDAT SERVICE SUITE - B2DROP



JÜLICH Forschungszentrum

B2DROP - B2SHAREBRIDGE PLUGIN

Use Case:

A researcher worked on a publication. The volatile digital objects were stored, synchronized and shared using B2DROP. Now, after finalizing it, he wants to publish the final document using B2SHARE.

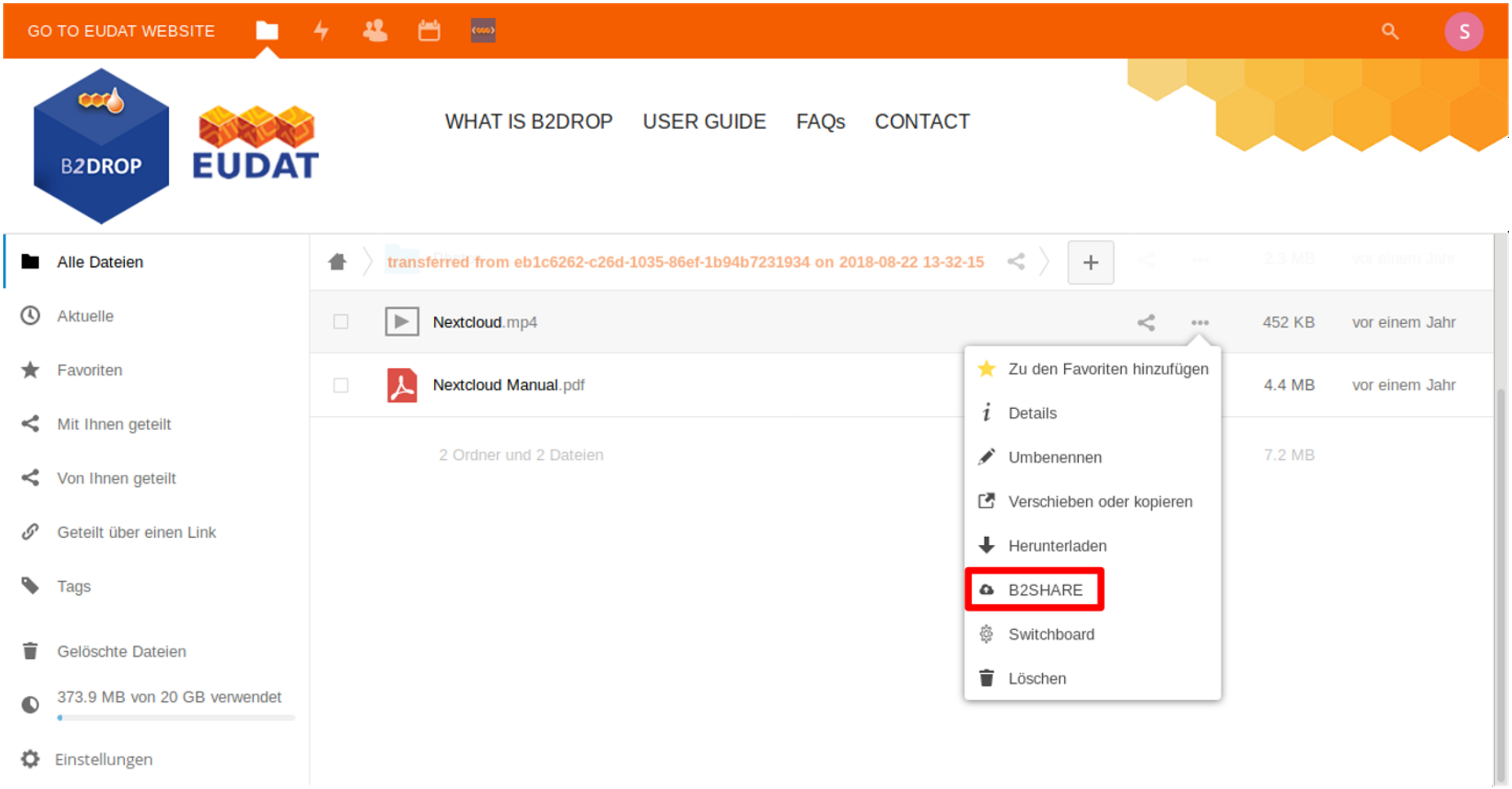
B2DROP - B2SHAREBRIDGE PLUGIN

Use Case:

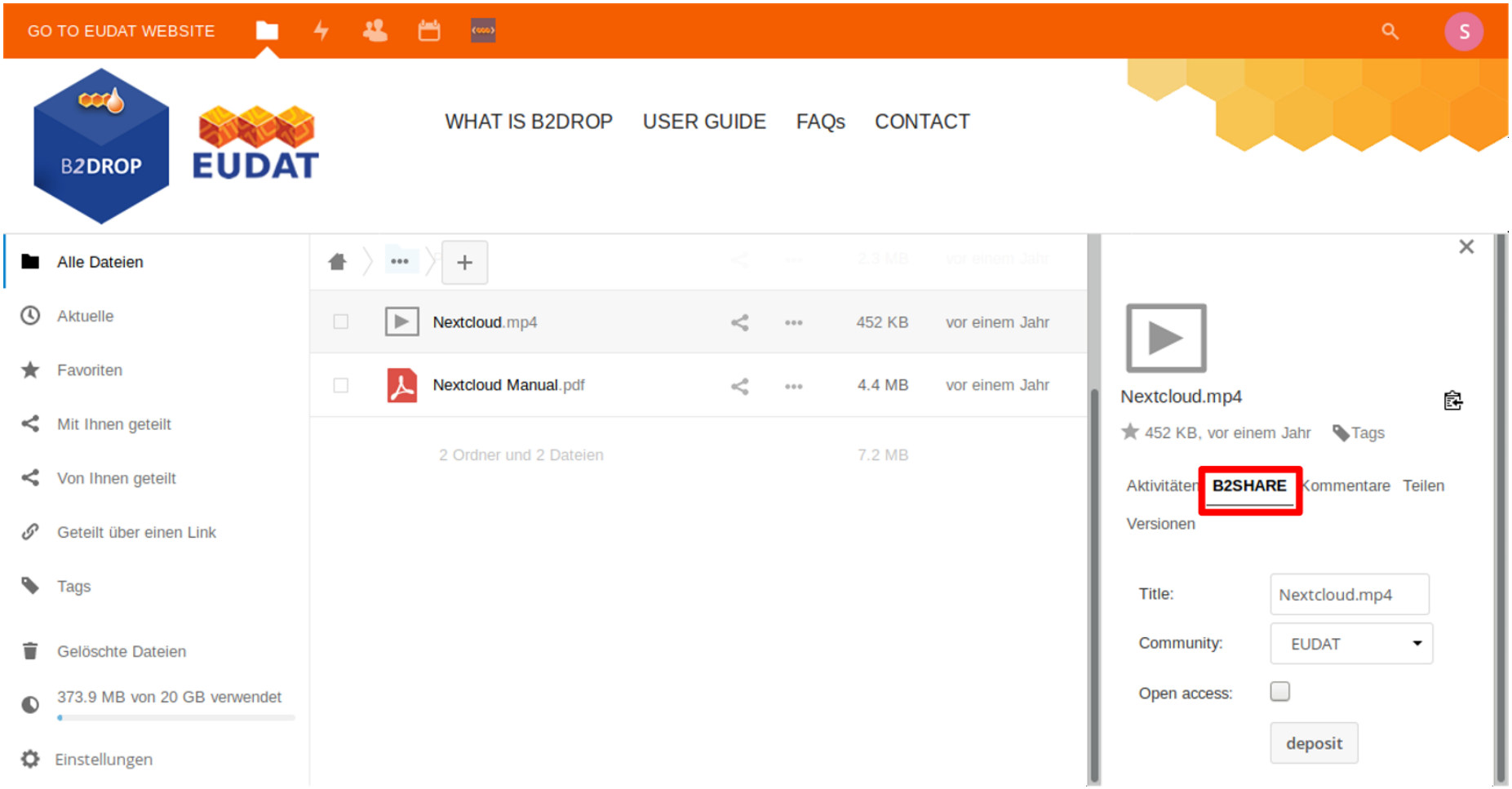
A researcher worked on a publication. The volatile digital objects were stored, synchronized and shared using B2DROP. Now, after finalizing it, he wants to publish the final document using B2SHARE.

- Extend B2DROP WebUI
  - Create deposit on a per-file base
  - Make community (metadata) selectable
- Implement third-party transfers

B2SHAREBRIDGE - FILE LIST

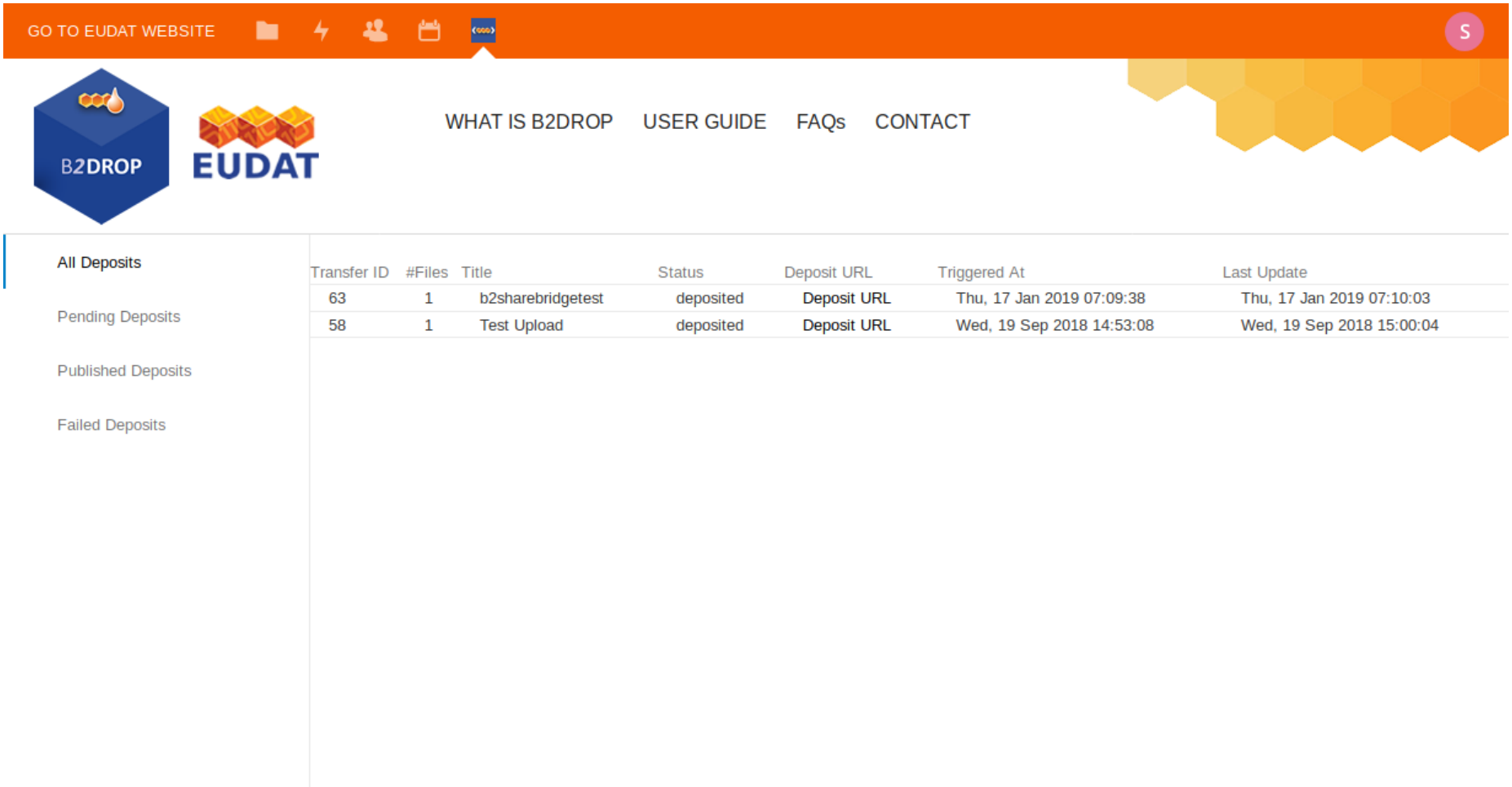


B2SHAREBRIDGE - COMMUNITY SELECTION

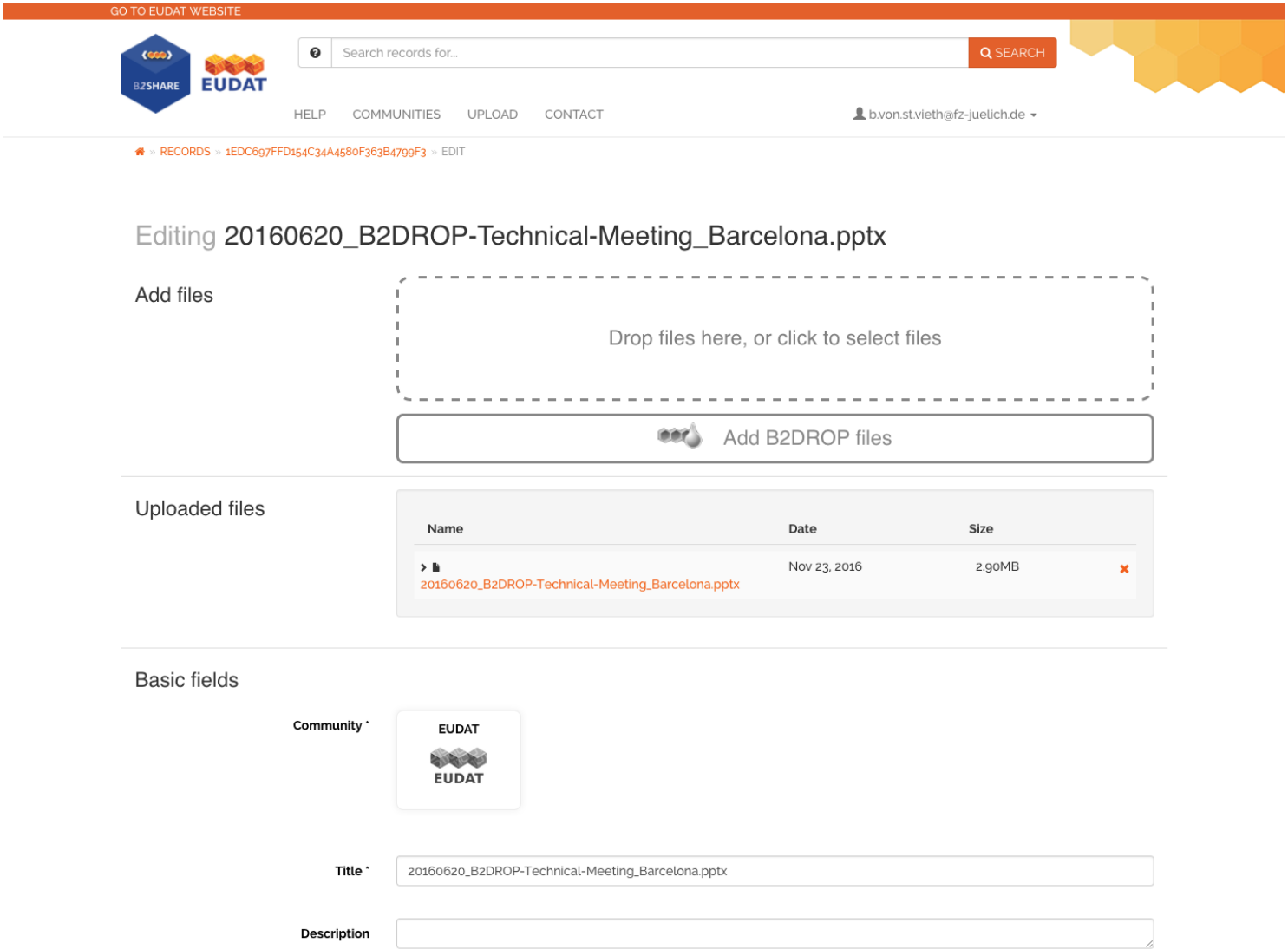


JÜLICH Forschungszentrum

B2SHAREBRIDGE - DEPOSIT LIST

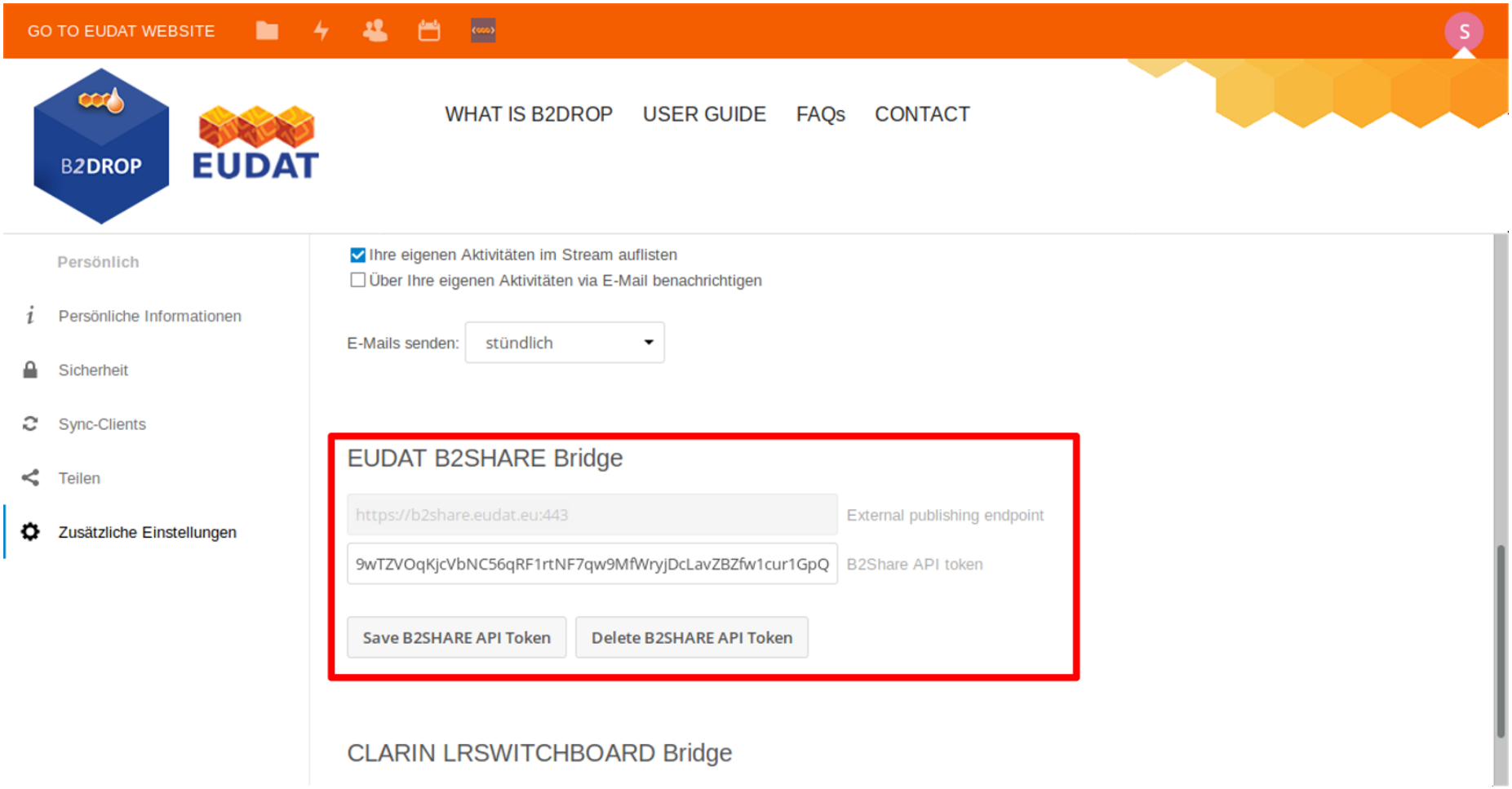


B2SHAREBRIDGE - DEPOSIT IN B2SHARE

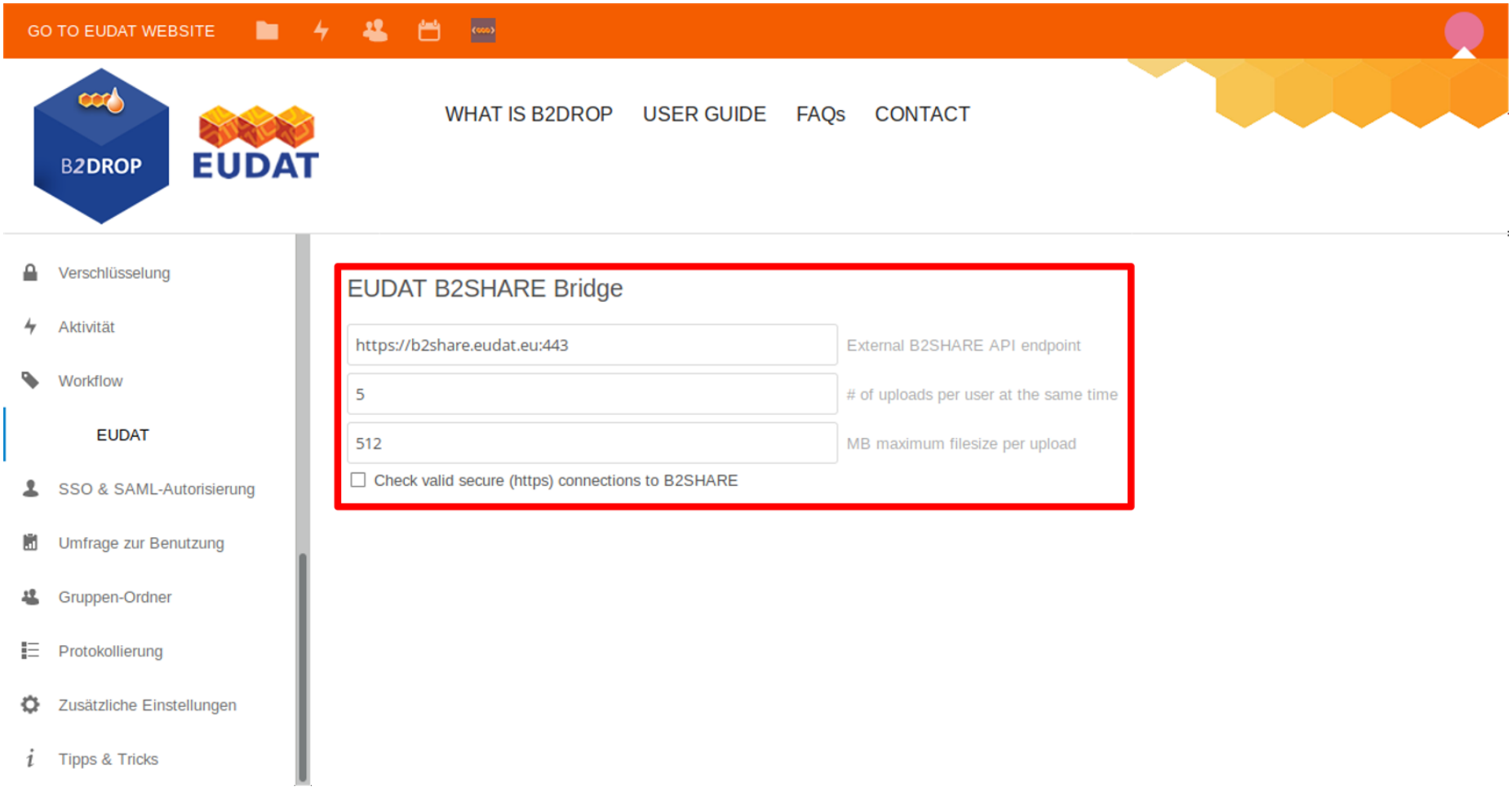


JÜLICH Forschungszentrum

B2SHAREBRIDGE - PERSONAL SETTINGS



B2SHAREBRIDGE - ADMIN SETTINGS



Hands-on session

PREPARATION

- Go to `b2drop-devel.zam.kfa-juelich.de/index.php`
- Select `B2ACCESS`
- Login with your existing account or create a new one
- Create a file with your user name from paper, e.g. haf13.txt
- Copy your cloud id from /settings/user/sharing in the created file
- Share the file with user sapweiler
- Create an app password at /settings/user/security and store it some where, you won't see it again

MOUNT YOUR B2DROP STORAGE

In some cases you want to mount your `B2DROP` storage locally:

- Use ssh to connect to `zam10141.zam.kfa-juelich.de`
- Create `.davfs2` folder in your home directory
- Create `~/.davfs2/secrets` file
- Write mountpoint, username and app password in the first line e.g.:`/home/haf13/b2drop sapweiler MYPASSWORD`
- Change permissions of secrets file to 600
- Run `mount b2drop` within your home directory

Prerequisite:

- Installation of davfs2
- User is member of davfs2 group
- Valid entry in `/etc/fstab`

USE YOUR B2DROP STORAGE

- Create a file with some content
- Move the file in the `B2DROP` folder
- Review if it is uploaded
- Share it with your neighbour
- Create a new txt file online
- Review if it is downloaded

PREPARE B2SHARE

- Go to `trng-b2share.eudat.eu`
- Login with your `B2ACCESS` account
- Create a new API token in your settings
- Store the API token in your personal B2DROP settings
- Store it in some file, it is needed later again

USE B2SHAREBRIDGE

Upload a file from B2DROP to B2SHARE using the B2SHAREBRIDGE

- Select the context menu of a file (· · ·)
- Select B2SHARE
- Enter a deposit name and select the metadata schema
- Press publish

The upload is done in background, during the next cronjob.

- Review you deposit status in the B2SHAREBRIDGE app
- If the status is deposited, click on the URL
- Finish your deposit
- Make another upload with two files

USE THE B2SHARE API

Using the B2SHARE depositer

- Fetch the Depositer from `B2DROP`
- View options and syntax `python3 run.py -h`
- Upload a file with the depositer, do not finalize it
- Review the draft online

USE THE B2SHARE API

Using the B2SHARE depositer

- Fetch the Depositer from `B2DROP`
- View options and syntax `python3 run.py -h`
- Upload a file with the depositer, do not finalize it
- Review the draft online

Using curl

- Identify a targeted community by listing the available ones
- Use the community identifier to fetch the community metadata schema
- Create a draft
- Upload files to your draft; one command per file
- Set the metadata to your record
- Publish your draft