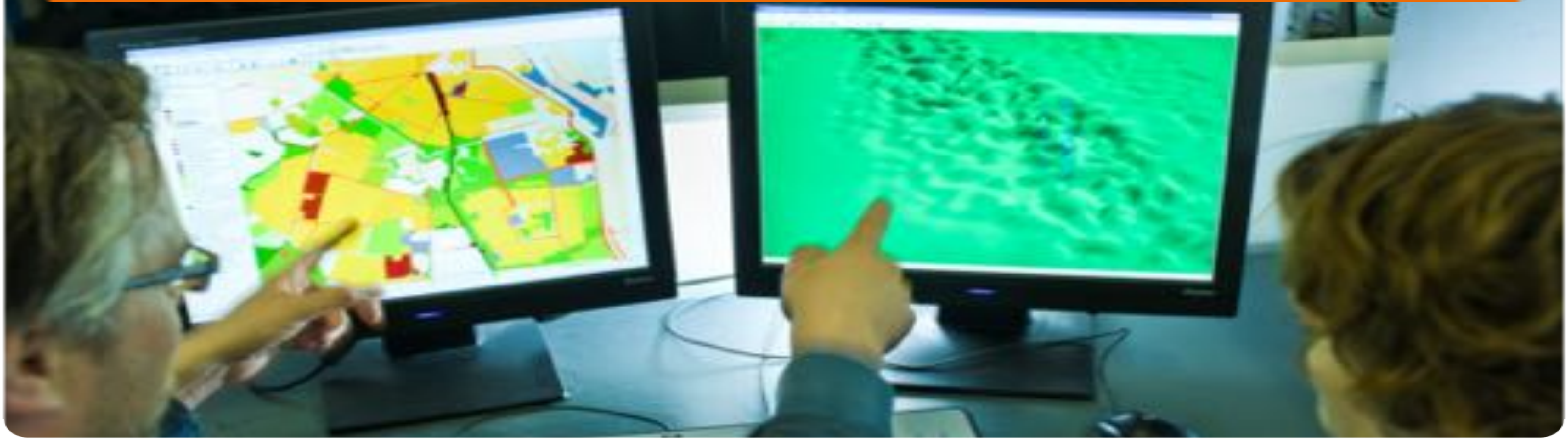
DATA MANAGEMENT IN COMPBIOMED MOVING TOWARDS FAIR DATA

NARGES ZARRABI



Machine Learning meets Modelling and Simulation Methods

17th March 2020

What is SURF?

SARA **(1971)** → SURFsara (2013) → **SURF (2020)**

SURF is the collaborative organisation for **ICT** in Dutch education and research

Driving innovation together

Fields of work

Education

Flexible education

Diverse learning resources

Using study data

Research

Unlimited access

World-class facilities

Stimulating Open Science

Cooperative facilities

On campus

Security in the digital world

User-centred

The Dutch National Supercomputer Cartesius

Total cores:	47767 CPU + 132 GPU
Total memory:	117 TB
Peak performance:	1.843 Pflop/sec
Disk space:	180 TB home, 7.7 PB project/scratch
Operating system:	BullX (GNU/Linux)
Network:	Mellanox InfiniBand 56 GBps bandwidth, 3µs latency

Top 500 largest supercomputers

- 2014: #45
- 2018: #360

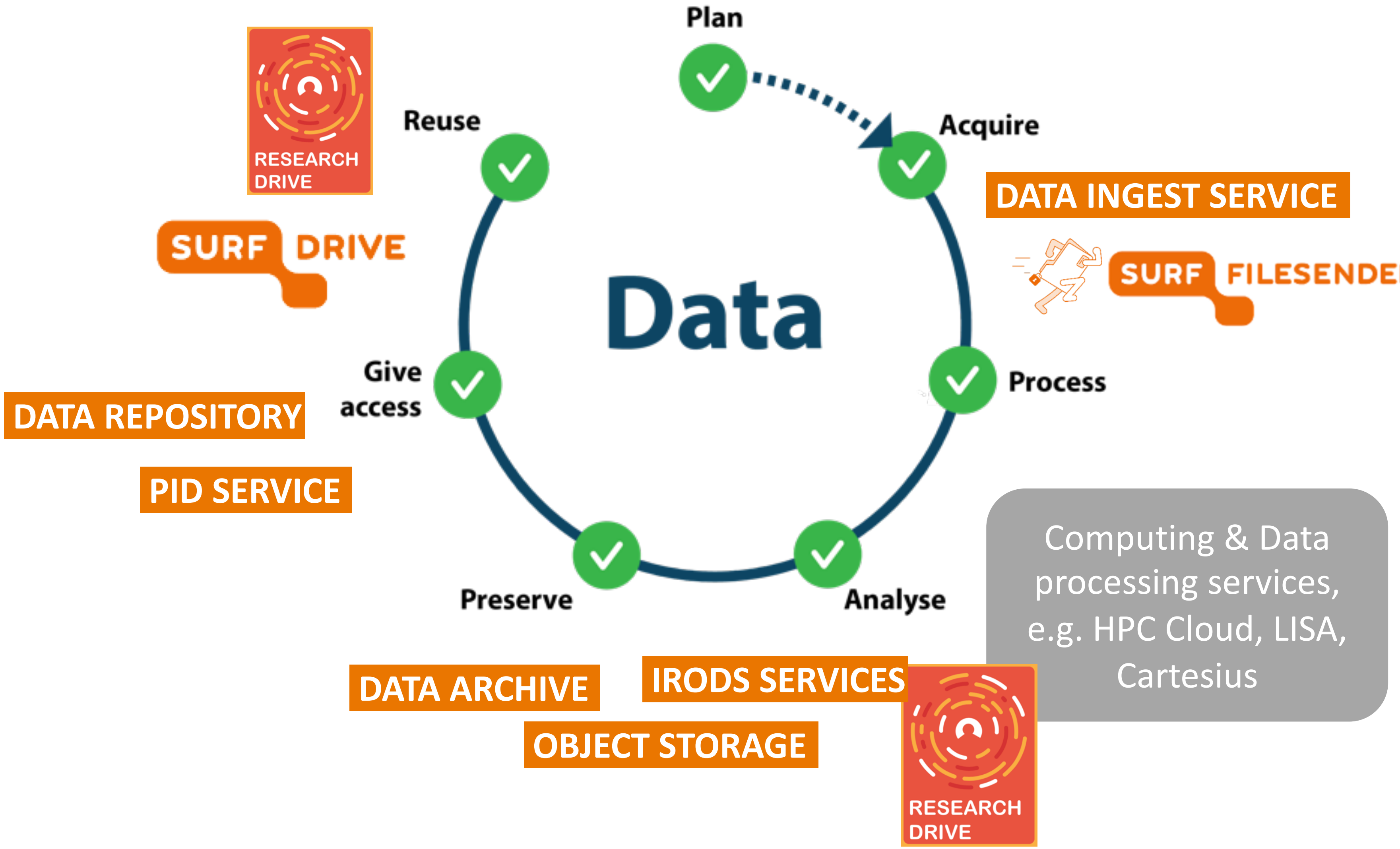
SURF is more..
than just big systems

Consultancy

Training

Knowledge Exchange

Data Services at SURF



Plan

Acquire

DATA INGEST SERVICE

SURF FILESENDER

Process

Computing & Data processing services, e.g. HPC Cloud, LISA, Cartesius

Analyse

IRODS SERVICES

RESEARCH DRIVE

OBJECT STORAGE

DATA ARCHIVE

Preserve

Give access

DATA REPOSITORY

PID SERVICE

Reuse

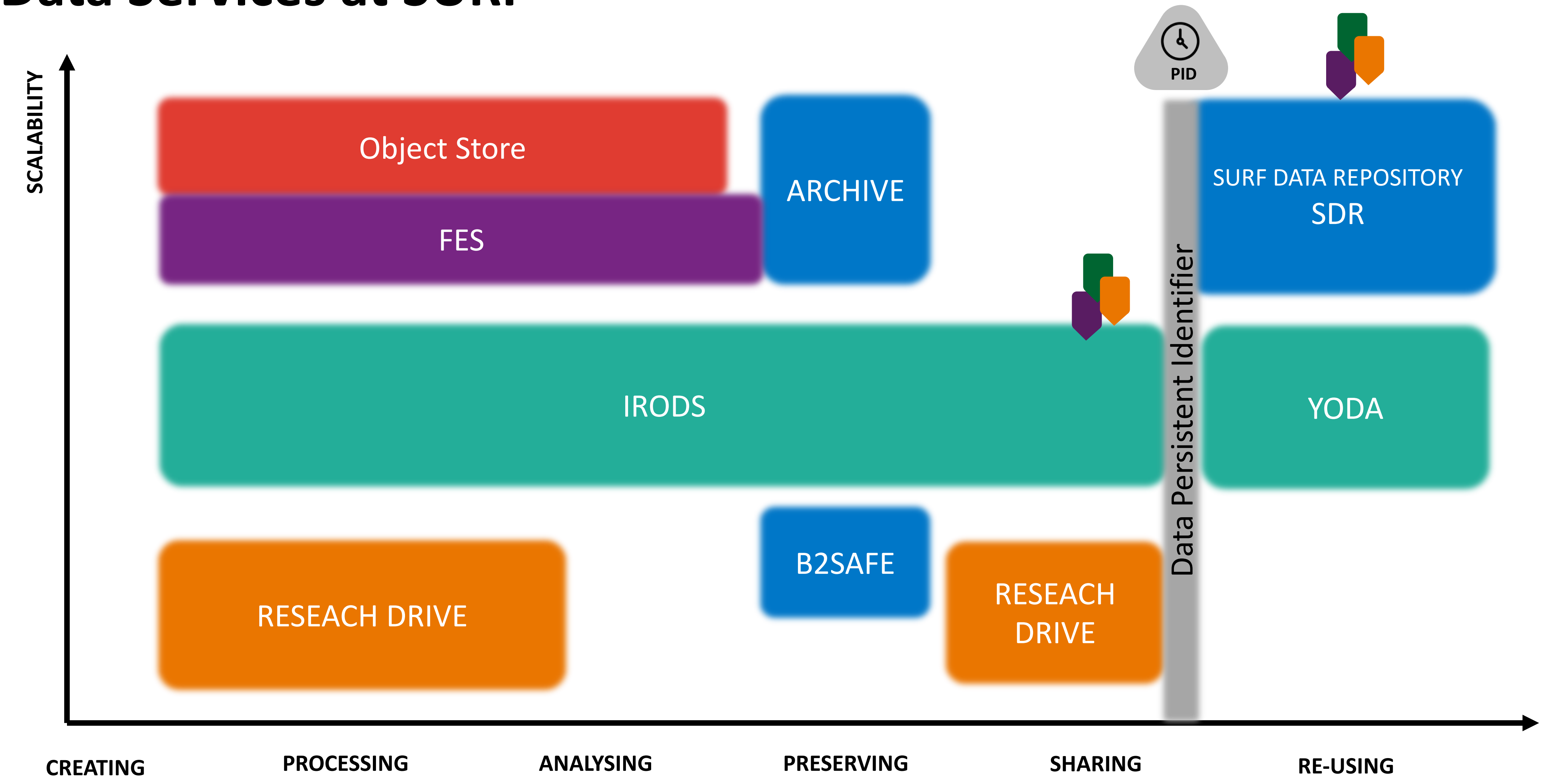
RESEARCH DRIVE

SURF DRIVE

Data

For more detail, please see https://www.surf.nl/diensten-en-producten/categorie/datadiensten

Data Services at SURF



Data Services Projects

EOSC-hub

FAIRsFAIR
Fostering Fair Data Practices in Europe

EUDAT Collaborative Data Infrastructure

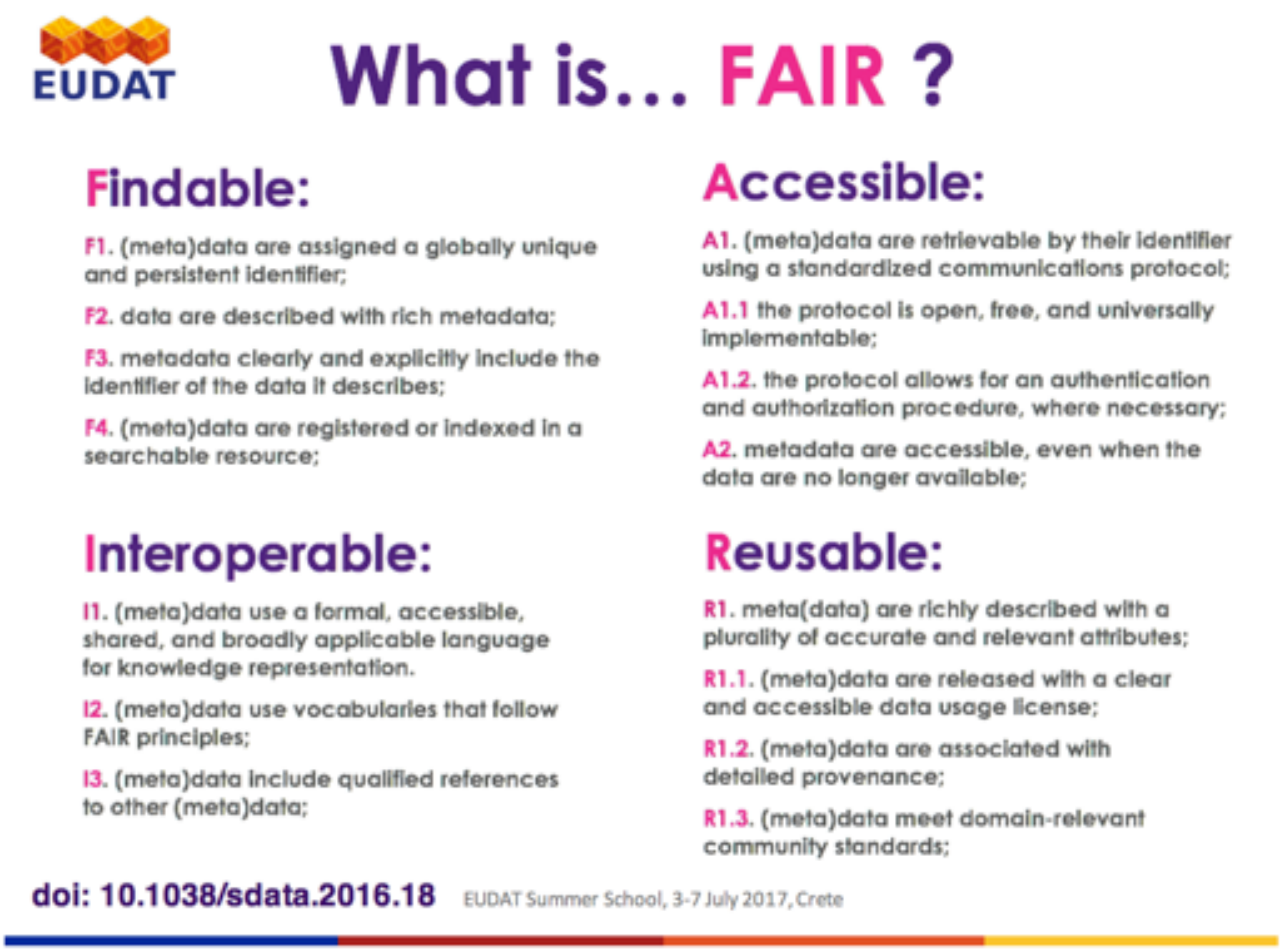
ENVRI FAIR

ODISSEI

CompBioMed

Data Requirements in Reseach Communities

- **More efficient data access, sharing and transfer**
  - *Intensive data-sharing and transfer*
  - *Restricted data-sharing and transfer*
- **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*



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;

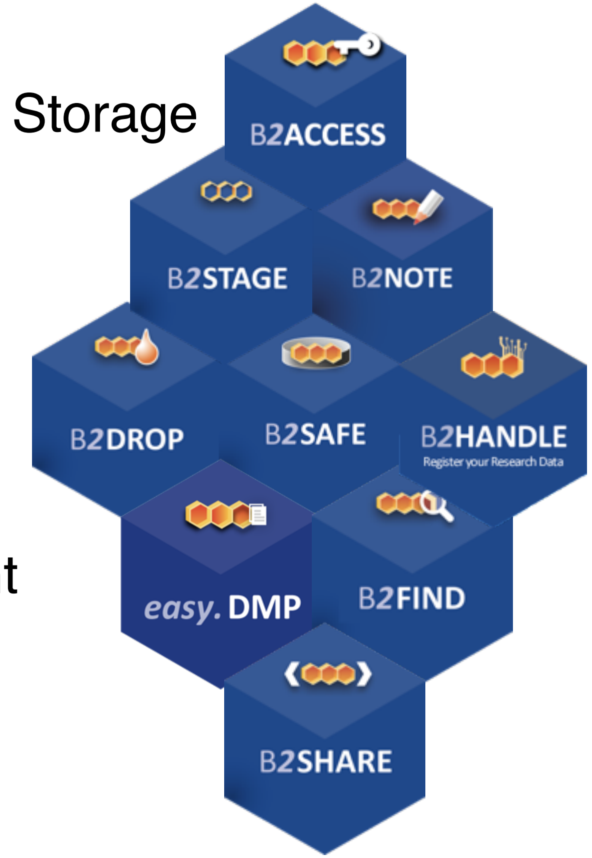
R1.3. (meta)data meet domain-relevant community standards;

doi: 10.1038/sdata.2016.18 EUDAT Summer School, 3-7 July 2017, Crete

EUDAT Collaborative Data Infrastructure

EUDAT B2Service Suite

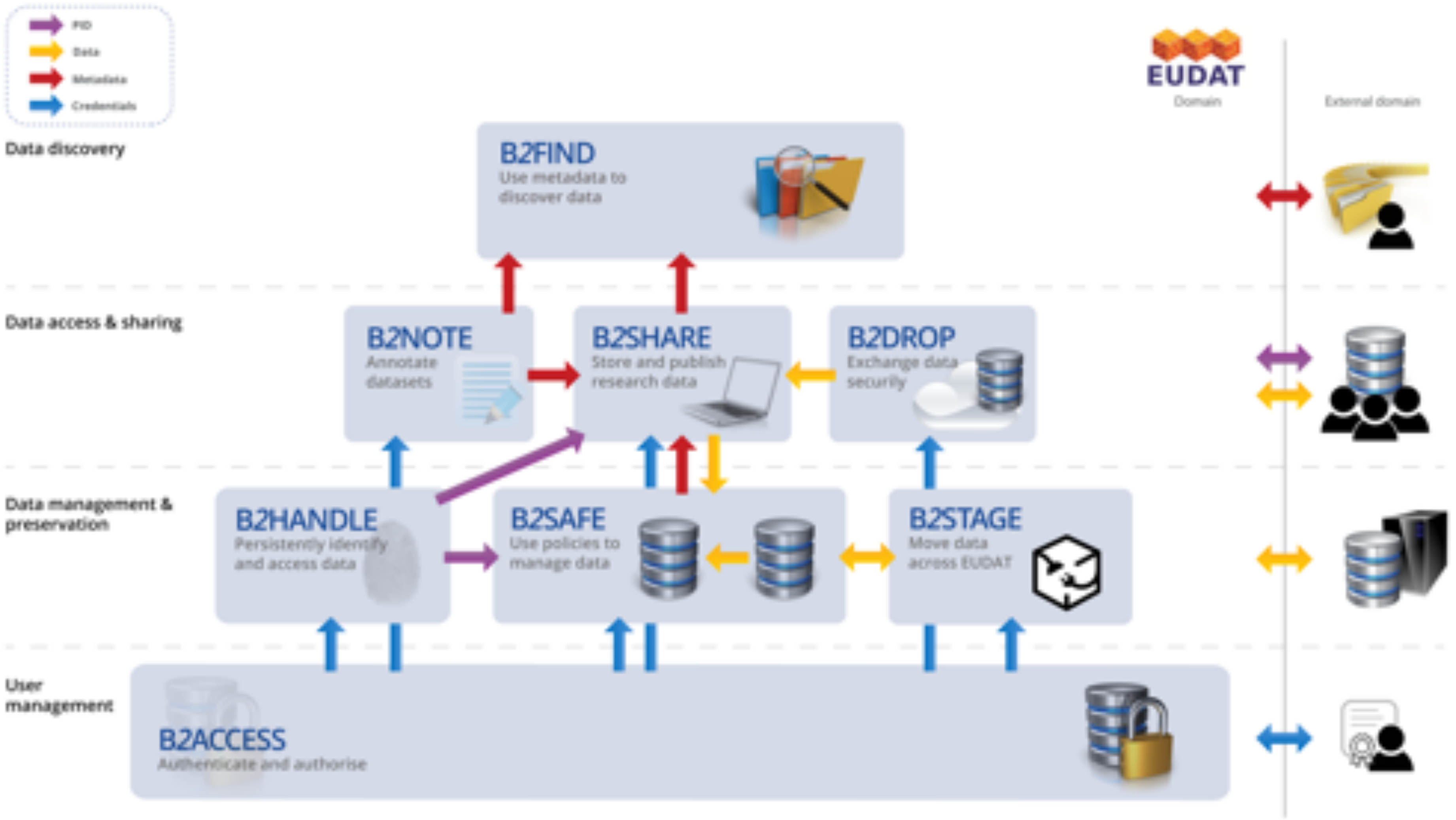
- **B2ACCESS** – Authentication and Authorisation
- **B2DROP** – Data Workspace
- **B2SAFE** – Distributed, Secure Policy Based Data Storage
- **B2SHARE** – Searchable Data Repository
- **B2STAGE** – High Performance Data Movement
- **B2FIND** – Searchable Metadata Aggregator
- **B2HANDLE** – Persistent Identifier Provider
- **B2NOTE** – Semantic Metadata Annotation
- **easy.DMP** – Data Management Planning Assistant
- **Gitlab** – Git repository and collaborative software development platform



SURF

EUDAT Collaborative Data Infrastructure

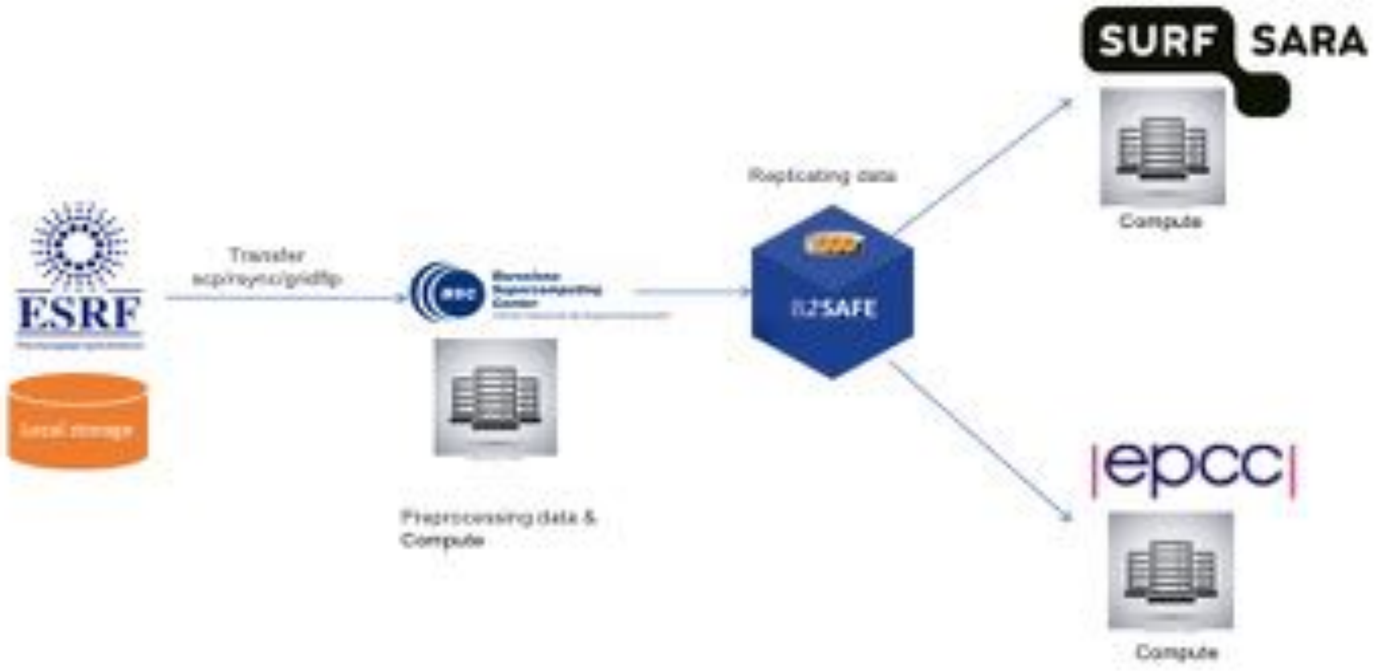
EUDAT B2 Services Diagram



SURF

Use case: Workflow using Alya Application

- **Step 1: Data creation and transfer:** The raw data is collected at ESRF in France. The data is being stored locally on tapes. Currently, a copy of the data is transferred to BSC.
- **Step 2: Data pre-processing:** In BSC, researchers pre-process the data which includes manual and automated steps for image stitching, segmentation and meshing.
- **Step 3: Data replication:** The preprocessed data needs to be replicated from BSC to SURFsara and EPCC. The replicated data will then be used to run simulations on the supercomputers in these sites.



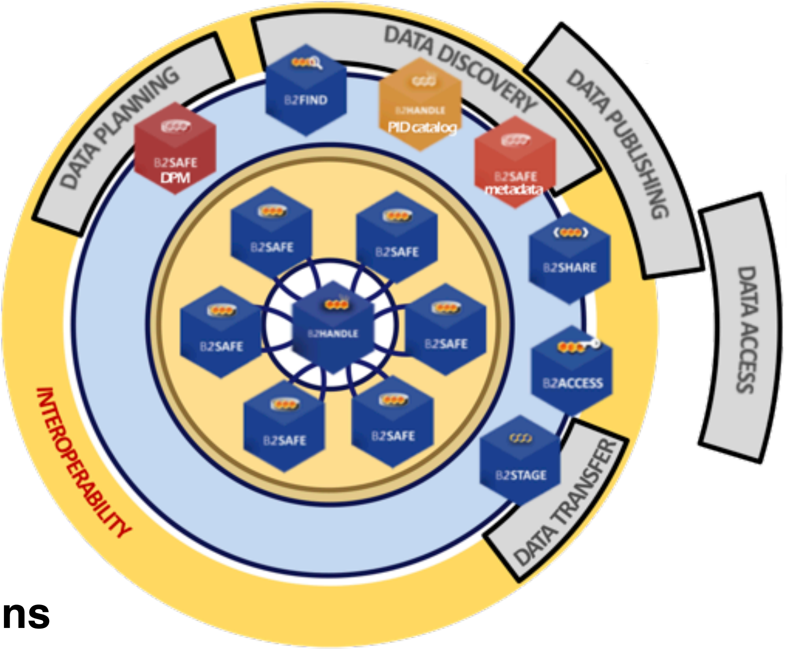
EUDAT Collaborative Data Infrastructure

B2SAFE

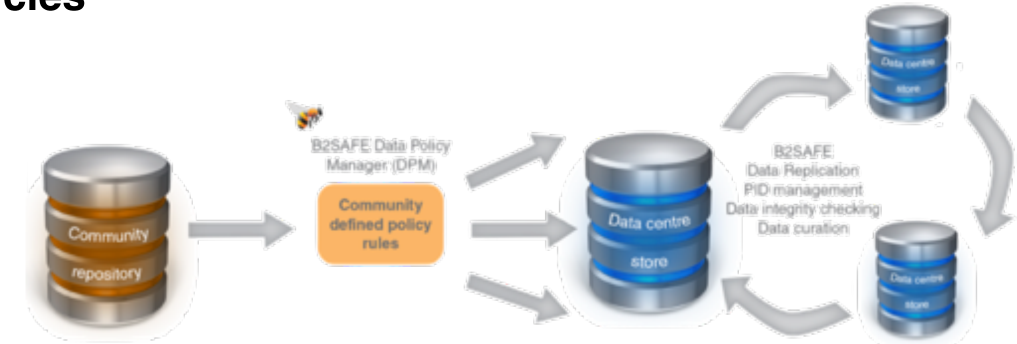
Replicate Research Data Safely

- **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

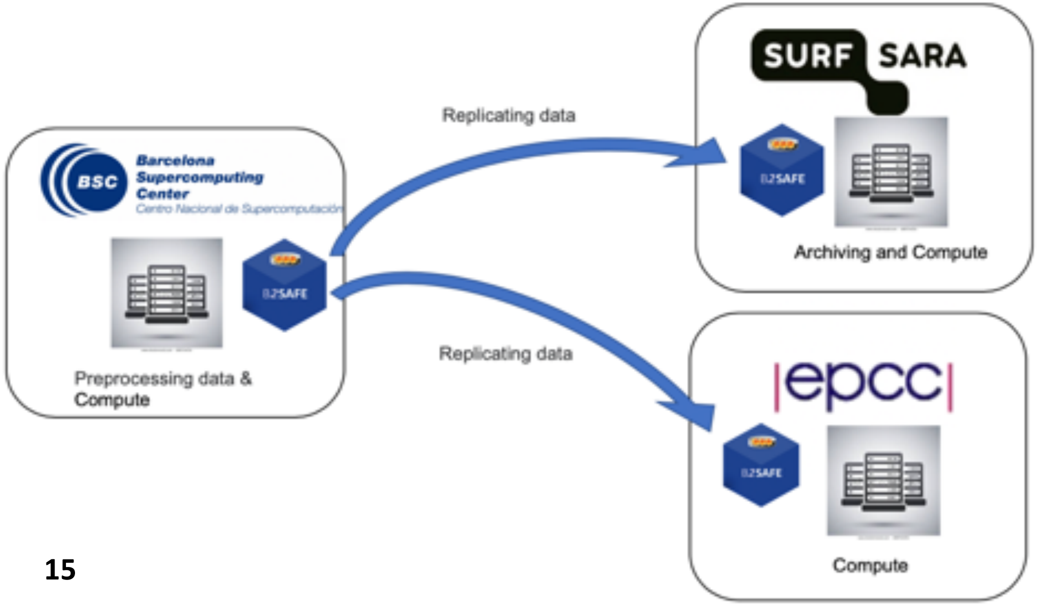


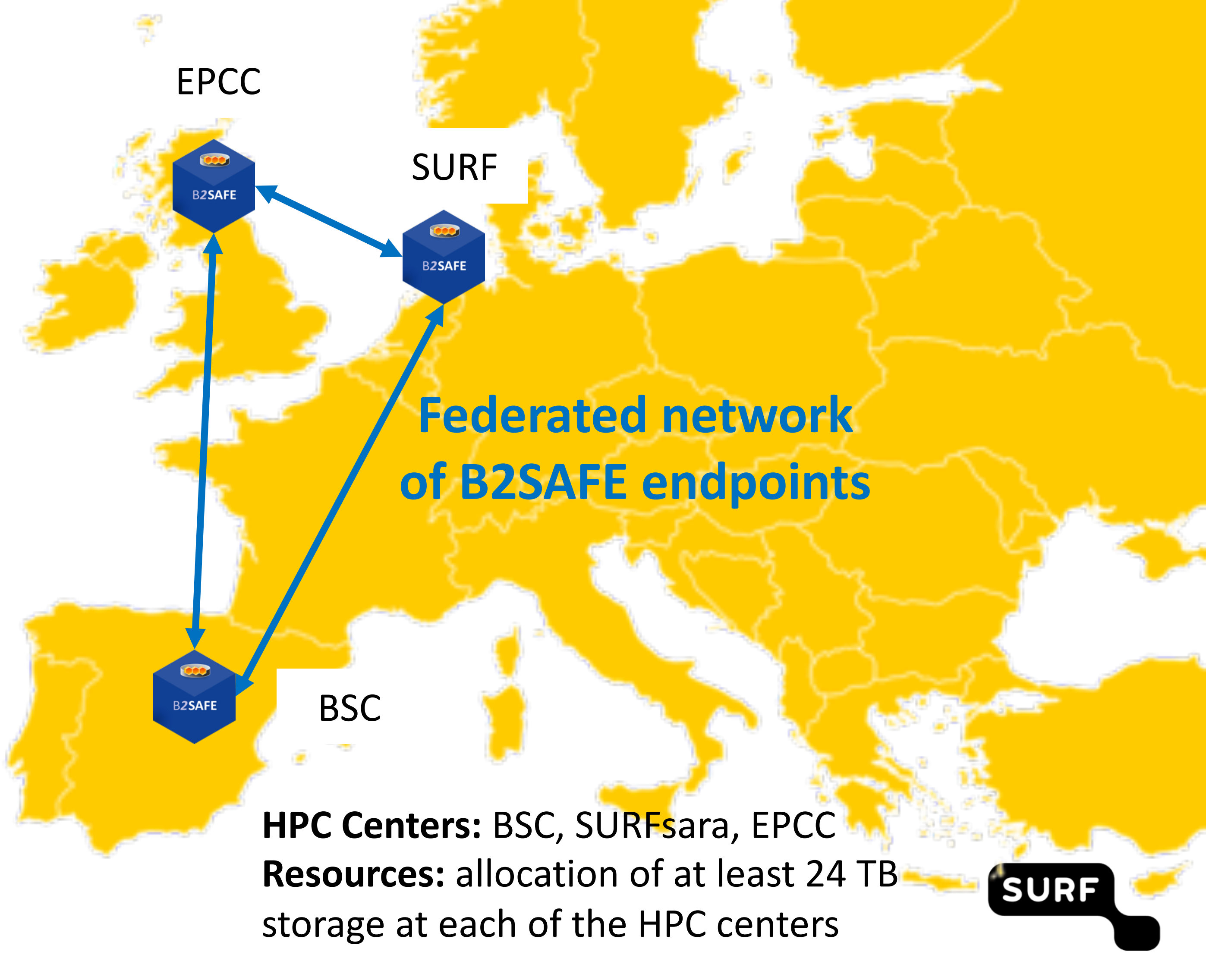
SURF

Data Replication Pilot (CompBioMed1)

Data replication pilot

- Safe data replication data preservation
- Allocation of PIDs to replicated data
- Facilitate large data transfer
- Bring data close to compute
- Scale-up compute power



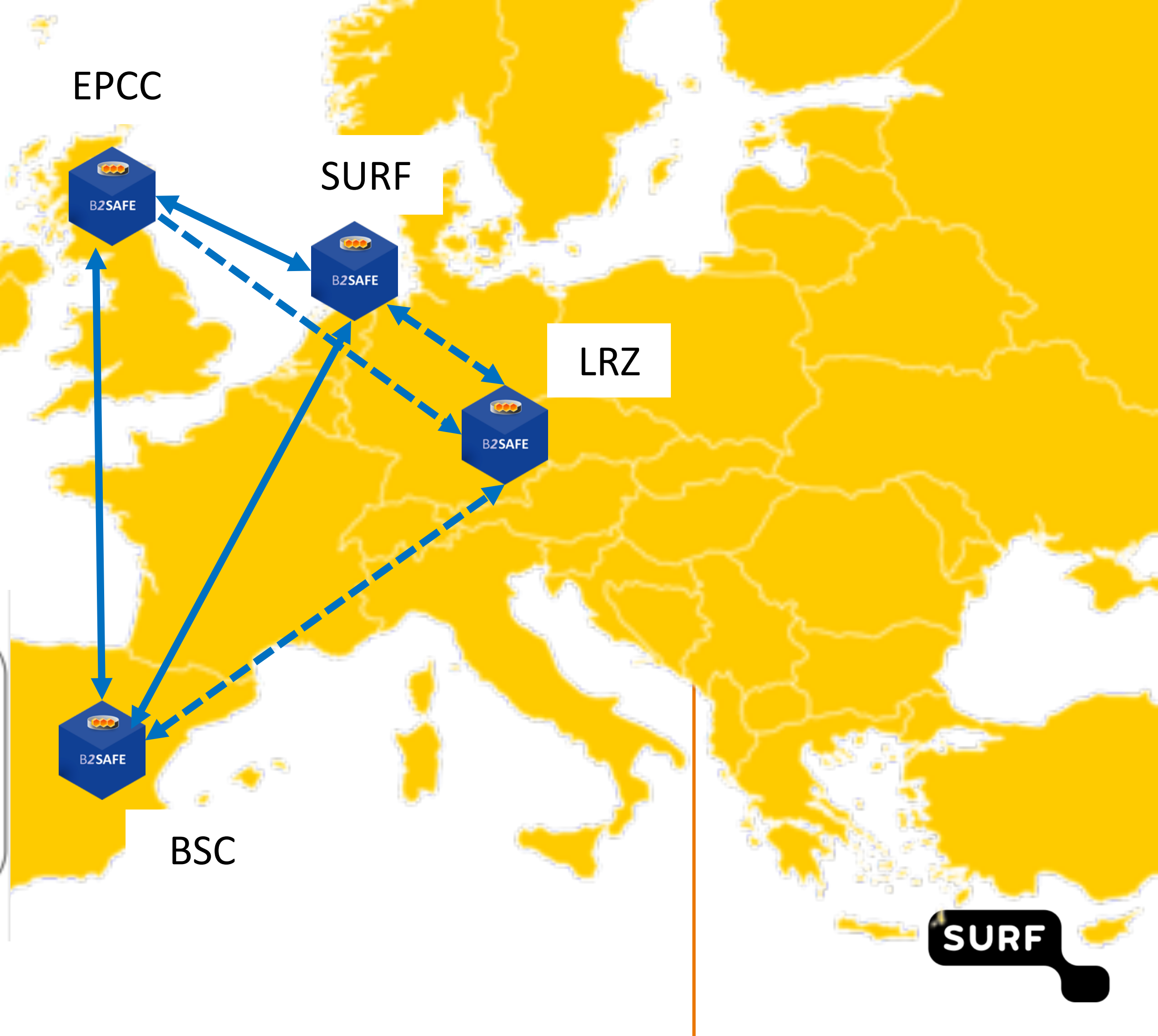


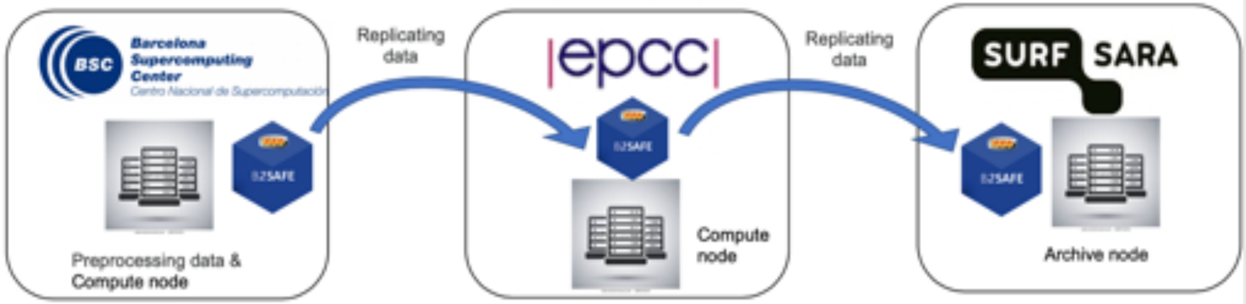
Federated network of B2SAFE endpoints

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

Future work pilot (CompBioMed2)

- Allocation of resources and replication of the real data (24 TB per Centre)
- Extend the network of B2SAFE endpoints including more HPC centres
- Try other replication scenarios
- Integration B2SAFE and B2SHARE

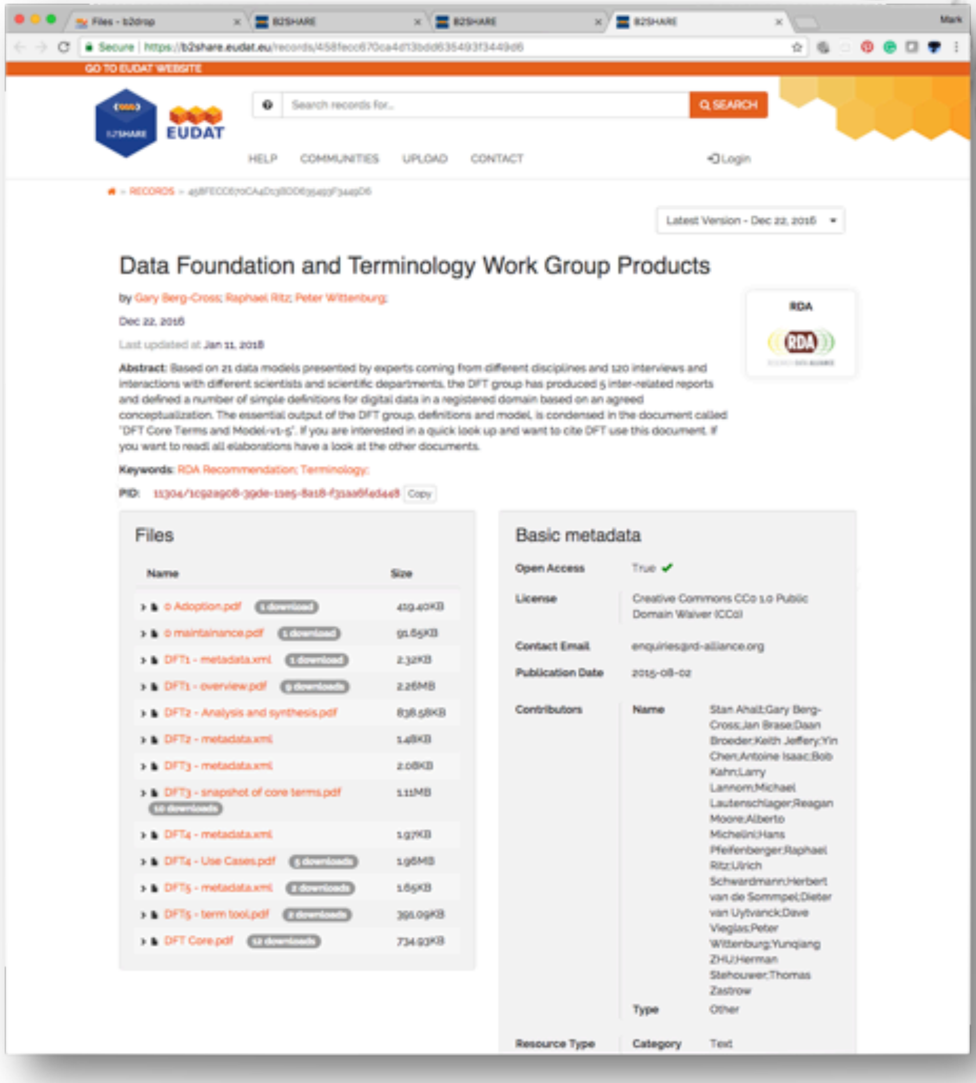



BSC
Barcelona Supercomputing Center
Centro Nacional de Supercomputación
B2SAFE
Preprocessing data & Compute node
Replicating data
epcc
B2SAFE
Compute node
Replicating data
SURF SARA
B2SAFE
Archive node

EUDAT Collaborative Data Infrastructure

B2SHARE
Store and Share Research Data

- **EUDAT Data Repository for publishing 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



Data Foundation and Terminology Work Group Products

by Gary Berg-Cross; Raphael Ritz; Peter Wittenburg;

Dec 22, 2016

Last updated at Jan 11, 2018

Abstract: Based on 21 data models presented by experts coming from different disciplines and 120 interviews and interactions with different scientists and scientific departments, the DFT group has produced 5 inter-related reports and defined a number of simple definitions for digital data in a registered domain based on an agreed conceptualization. The essential output of the DFT group, definitions and model, is condensed in the document called "DFT Core Terms and Model-v1-5". If you are interested in a quick look up and want to cite DFT use this document. If you want to read all elaborations have a look at the other documents.

Keywords: RDA Recommendation; Terminology;

PID: 11304/1c92a908-39de-11e5-8a18-f31aa6f4d448 Copy

Files

Name	Size
0 Adoption.pdf 1 download	419.40KB
0 maintainance.pdf 1 download	91.65KB
DFT1 - metadata.xml 1 download	2.32KB
DFT1 - overview.pdf 9 downloads	2.26MB
DFT2 - Analysis and synthesis.pdf	838.58KB
DFT2 - metadata.xml	1.48KB
DFT3 - metadata.xml	2.08KB
DFT3 - snapshot of core terms.pdf 10 downloads	1.11MB
DFT4 - metadata.xml	1.97KB
DFT4 - Use Cases.pdf 5 downloads	1.98MB
DFT5 - metadata.xml 2 downloads	1.65KB
DFT5 - term tool.pdf 2 downloads	391.09KB
DFT Core.pdf 12 downloads	734.93KB

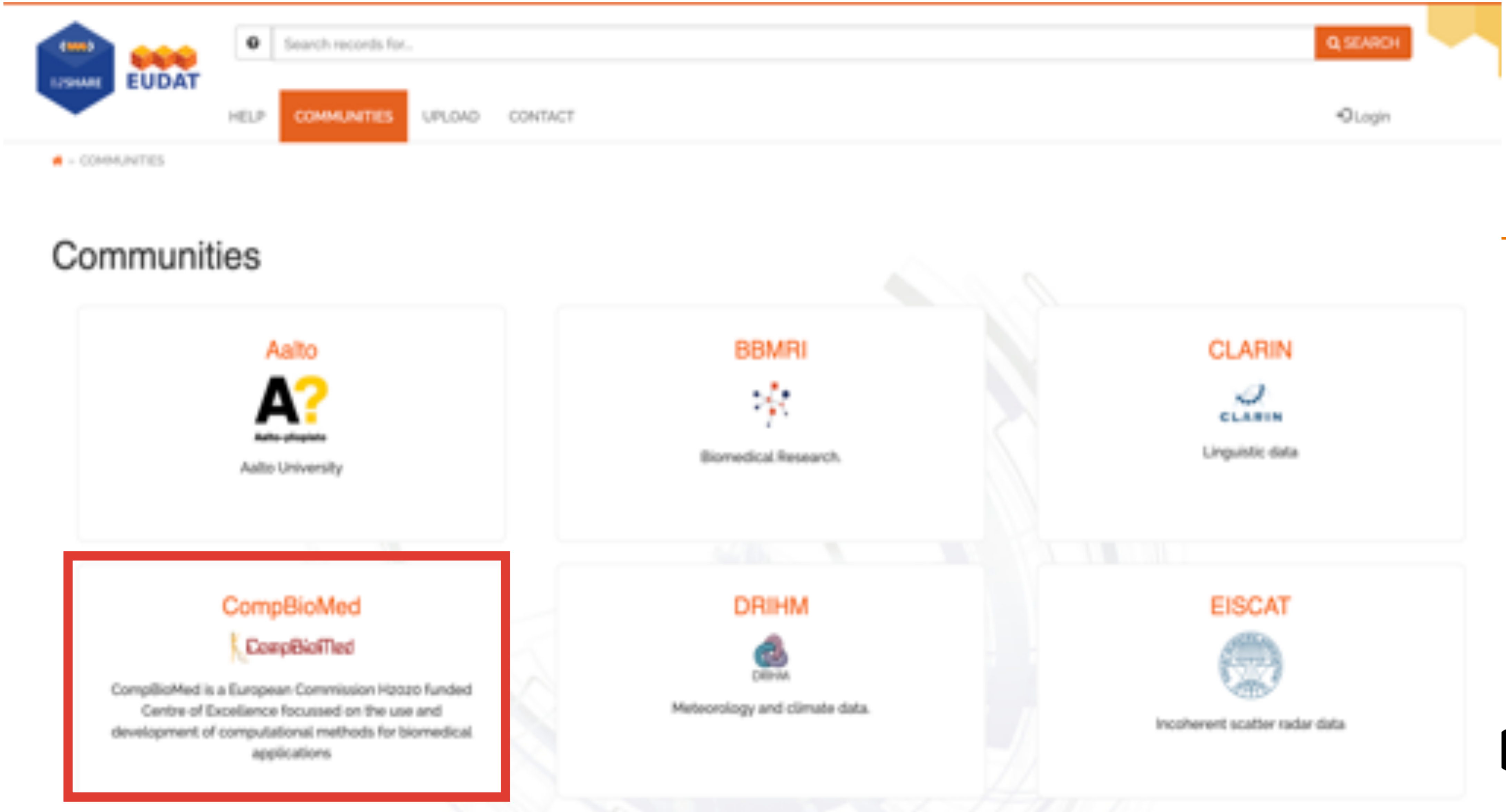
Basic metadata

Open Access	True ✓	
License	Creative Commons CC0 1.0 Public Domain Waiver (CC0)	
Contact Email	enquiries@rd-alliance.org	
Publication Date	2015-08-02	
Contributors	Name	Stan Ahalt;Gary Berg-Cross;Jan Brase;Daan Broeder;Keith Jeffery;Yin Chen;Antoine Isaac;Bob Kahn;Larry Lannom;Michael Lautenschlager;Reagan Moore;Alberto Michelini;Hans Pfeifenberger;Raphael Ritz;Ulrich Schwardmann;Herbert van de Sommpel;Dieter van Uytvanck;Dave Vieglas;Peter Wittenburg;Yunqiang ZHU;Herman Stehouwer;Thomas Zastrow
	Type	Other
Resource Type	Category	Text

SURF

https://b2share.eudat.eu/

CompBioMed Community in B2SHARE



Communities

Aalto
Aalto University

BBMRI
Biomedical Research.

CLARIN
Linguistic data

CompBioMed
CompBioMed is a European Commission H2020 funded Centre of Excellence focussed on the use and development of computational methods for biomedical applications

DRIHM
Meteorology and climate data.

EISCAT
Incoherent scatter radar data

Thank you!

Thanks to:

- SURF Data Preservation Services team
- SURF Data Management Services team
- Projects:

CompBioMed

EUDAT Collaborative Data Infrastructure

EOSC-hub