DATA MANAGEMENT IN COMPBIOMED
Moving towards FAIR data using DICE services

DICE COMPBIOMED ROADSHOW WEBINAR - 29TH OCTOBER 2021

Narges Zarrabi

Senior Advisor, SURF

SURF is the collaborative organisation for IT

in Dutch education and research

DiCE Data Infrastructure Capacity for EOSC

EOSC-hub

FAIRSFAIR Fostering Fair Data Practices in Europe

EUDAT Collaborative Data Infrastructure

CompBioMed

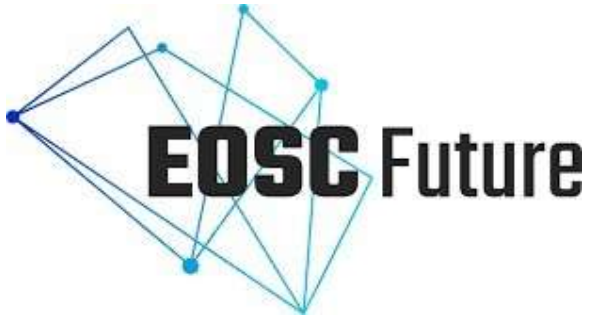
EOSC Future

Consultancy

Training

Knowledge Exchange

SURF



Projects at SURF

What do I do?

SURF

- Senior Advisor in Research Data Management at SURF since 2016

CompBioMed:

- Involved in WP3 on Data Management and Analytics as a consultant, since 2017

DICE

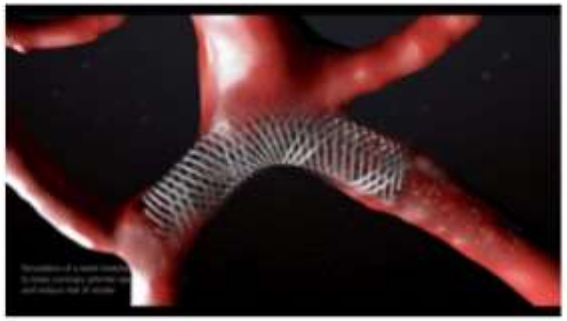
- Involved in WP5 on integration with community platforms
- Leading Task 5.2: CompBioMed Data Platform integration

Computational BioMedicine

Academic Users

In this section you will find links relevant to Academic Users including user case studies, and information from our Academic Partners.

Industrial Users



In this section you will find links relevant to Industrial Users including user case studies, and information from our Industrial Partners.

Clinical Users

In this section you will find links relevant to Clinical Users including user case studies, and information from our Partners working with medical institutions.

General Public

For those from the general public and media who are interested in our project and what we are planning follow this link and the relevant links on the page.

Data Management Challenges of Research Communninties

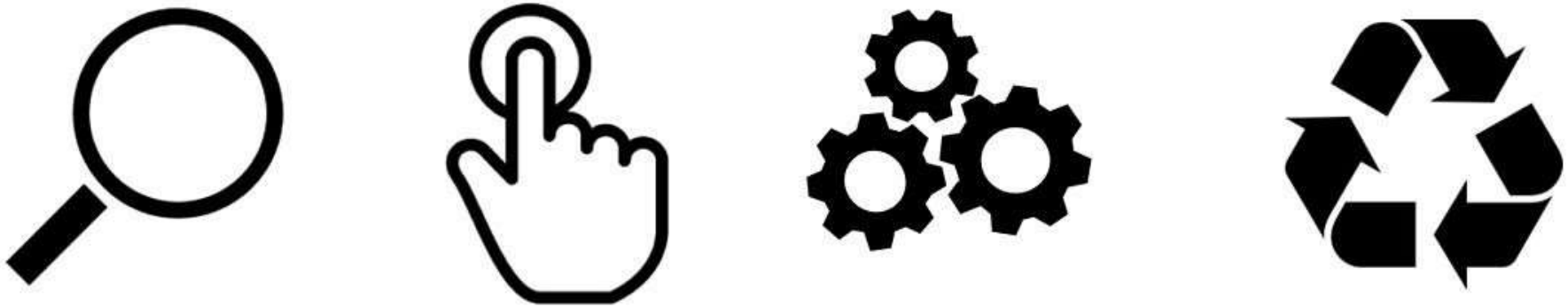
- **More efficient data access, sharing and transfer**
  - *Intensive data-sharing and transfer*
  - *Restricted data-sharing and transfer*
- **Preserving research data**
  - *Storage, backup and archiving large data, synchronizing data over distributed places*
  - *data provenance*
- **Accessible research 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

Main Challenge to make data FAIR

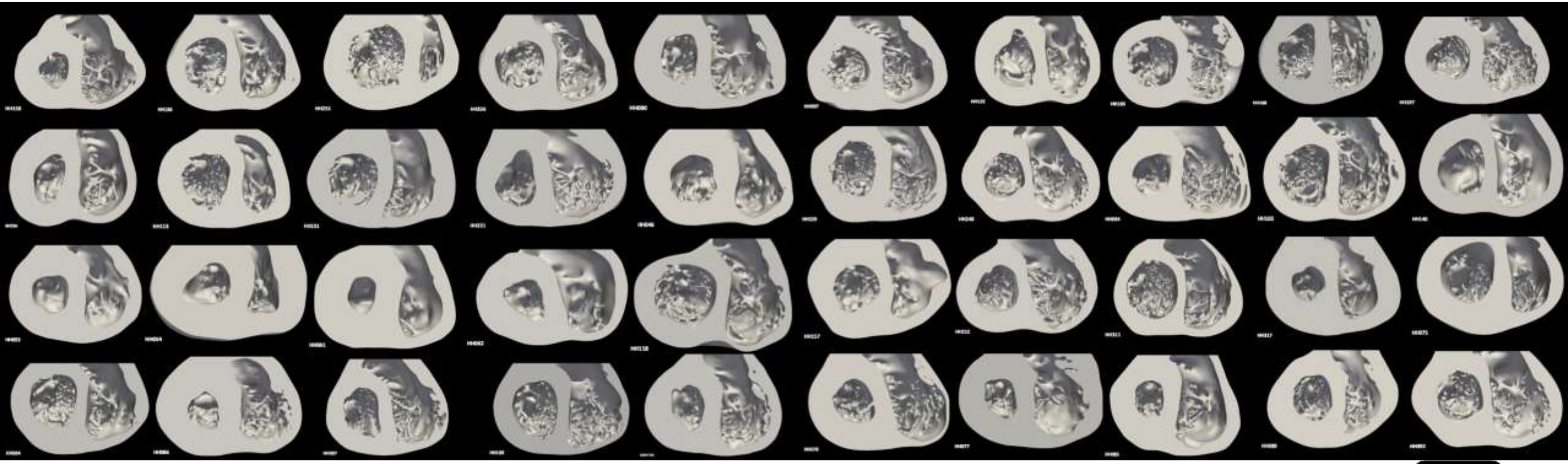
- Lack of an encompassing solution for publishing data and/or metadata
- Technical knowledge and awareness for producing FAIR data

Findable Accessible Interoperable Reusable



Example research use case with Alya application

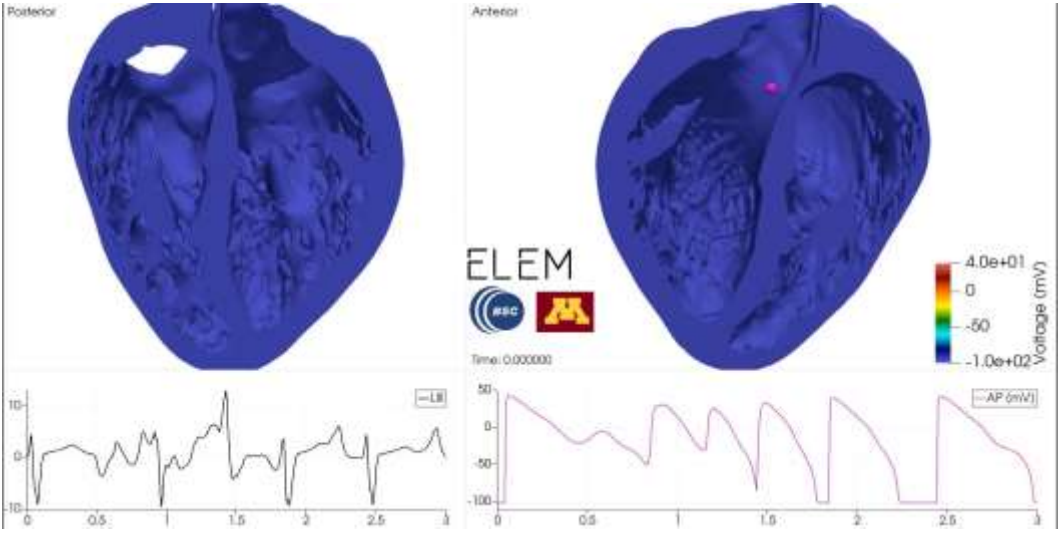
In-Silico Human Clinical Trial for Cardiac Safety Assessment of Drugs



- Why do drugs may produce pro-arrhythmic effects on some people and not others?
- Can we reproduce this observed behaviour to create a normal human in-silico population?

In-Silico Trial was able to reproduce the potential arrhythmic effect of the combined use of Hydroxychloroquine and Azithromycin

- An in-silico trial to assess 7 drug doses and drug combinations yielded 27Tb of data.
- Not only ion channel kinetics are determinant of drug-induced arrhythmic risk, but ANATOMY itself influences that risk.

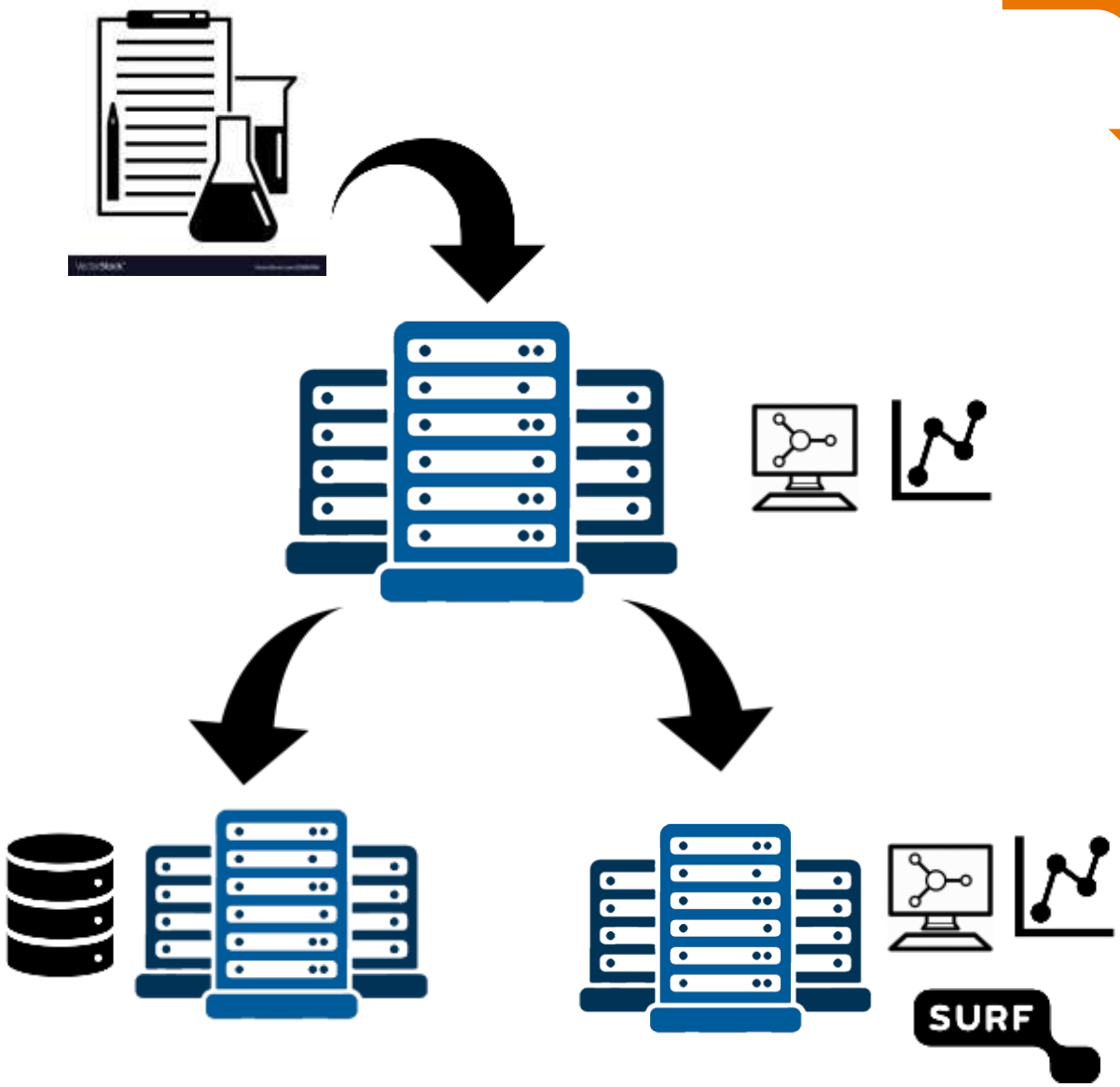


	Percentage (%) Virtual Population (N=64)	Clinical Data (N=90) Mercuro et al.	Clinical Data (N=200) Saleh et al.
Baseline	0	NA	NA
Hydroxychloroquine 800 mg	43.7%	NA	NA
Hydroxychloroquine 400 mg	**21.8%**	**19%**	NA
Hydroxychloroquine 200 mg + Azithromycin 500 mg	**9.3%**	NA	**3.5%**
Hydroxychloroquine 400mg + Azithromycin 500mg	**21.8%**	**21%**	NA
Azithromycin 500 mg	1.5%	NA	NA
Hypokalaemia (3.2 mol K)	20.3%	NA	NA
Hypokalaemia, Hydroxychloroquine 400mg + Azithromycin 500mg	64%	NA	NA

Mercuro JN et al. Risk of QT interval prolongation associated with the use of hydroxychloroquine with or without concomitant azithromycin among hospitalised patients testing positive for coronavirus disease 2019 (COVID-19). JAMA Cardiol. 2020; 5(9):1036-1041.
Saleh et a. Effect of Chloroquine, Hydroxychloroquine, and Azithromycin on the corrected QT interval in patients with SARS-CoV-2 infection. Circ. Arrhythm Electrophysiol. 2020 Jun; 13(6):e008662.

Workflow using Alya Application

- **Step 1: Data creation and transfer:** The raw data is collected at a lab (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 other HPC



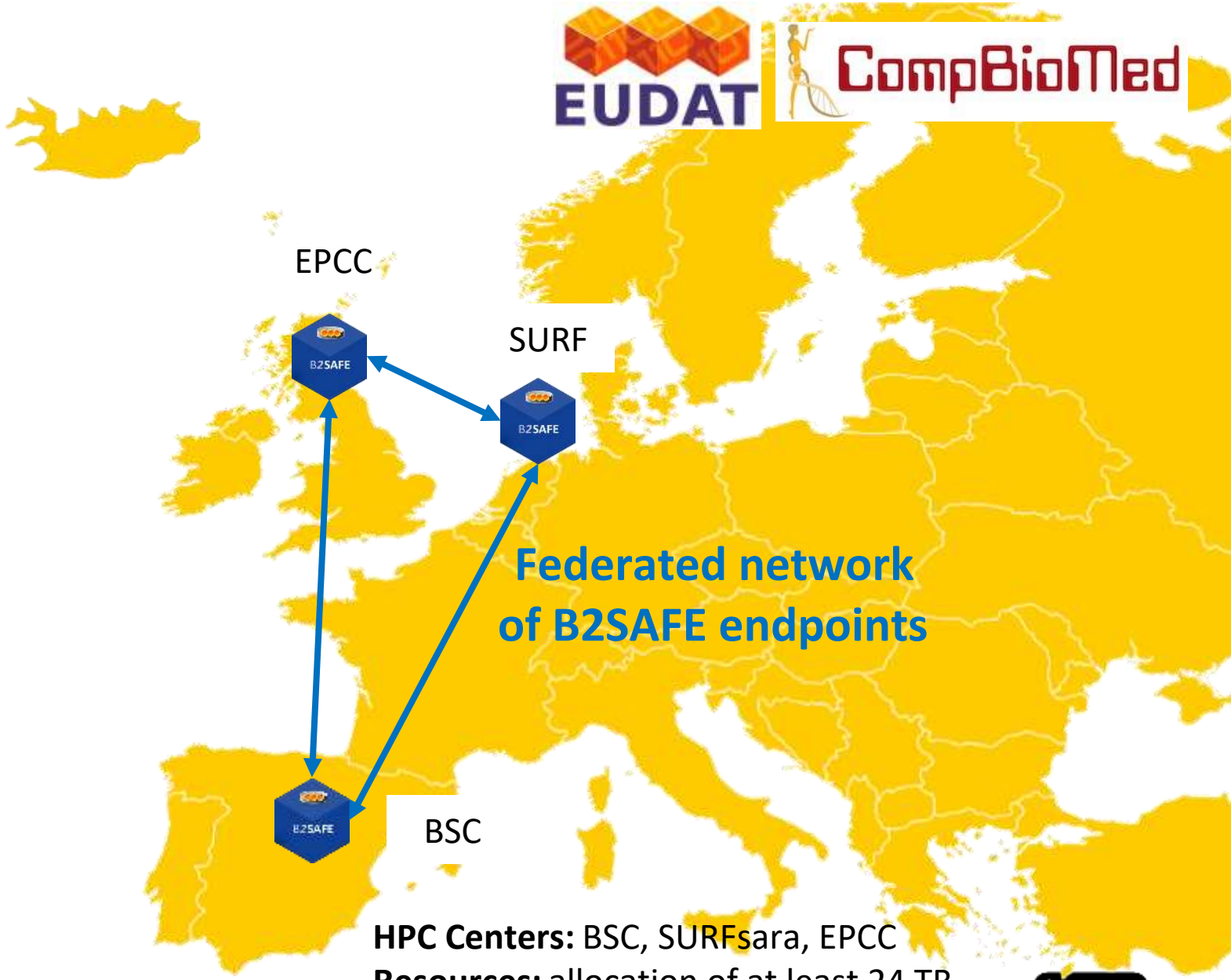
centers such as SURF. The replicated data will then be used to run simulations on the supercomputers in these sites.

- **Step 4: Data Processing and analysis:** running simulations and analyze output data

Data Replication Pilot (CompBioMed1)

Aim

- Safe data replication and data preservation
- Facilitate large data transfer
- Bring data close to compute
- Scale-up compute power **Achieved**:
- Deployed B2SAFE in BSC and EPCC
- Federated the 3 HPC centers (SURF, BSC, EPCC)



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

- Tested the federation and transfer performance

CompBioMed and DICE collaboration

- **Workflows to be implemented:**
  - **Data replication workflow: facilitate large data transfer by making replicas, data preservation, bring data close to compute**
  - **Data publication workflow: A data repository for publishing (large) data and/or metadata, metadata schema for CompBioMed**
- **CompBioMed partners involved: UCL, BSC, SURF**
- **EUDAT and DICE services to be used:**

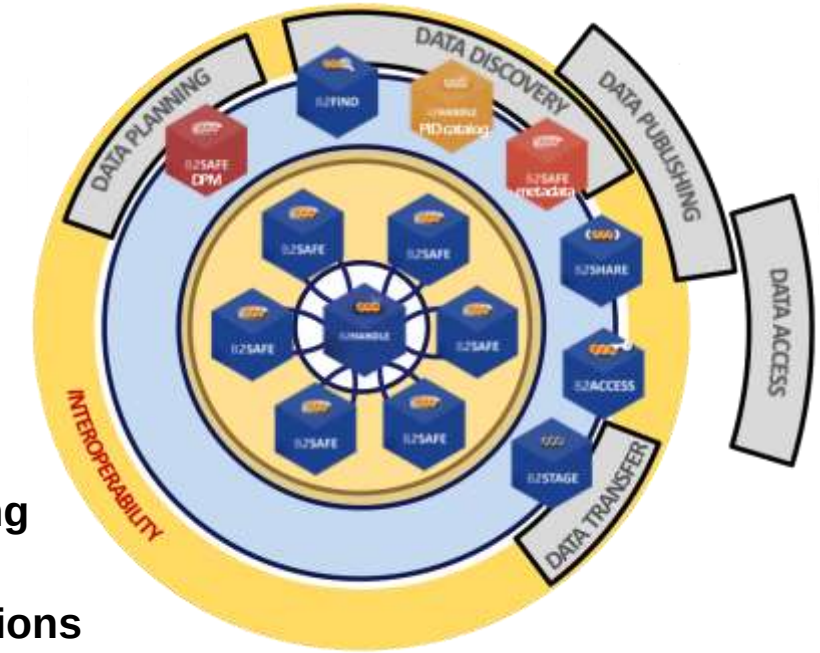
- **B2SHARE** – Searchable Data Repository
- **B2HANDLE** – Persistent Identifier Provider
- **B2SAFE** – Distributed, Secure Policy Based Data Storage

DICE services for CompBioMed

Service	*Description*	*Resources Needed*	*Provider*
B2SHARE	Data Repository for data publication. Metadata schema can be implemented in this repository. Integration with B2FIND for harvesting data and facilitating findability of the data.	50 TB	UCL
B2HANDLE	Tool required to make persistent identifiers (PIDs) for the data to facilitate findability of the data. The PIDs will potentially be used in B2SAFE and B2SHARE.	1 prefix 10000 PIDs	SURF
B2SAFE	Data staging and safe replication of research data between HPC centers in CompBioMed. The archival storage on tape facilitates long-term preservation of the data.	50 TB 50 TB	SURF BSC

EUDAT Collaborative Data Infrastructure

B2SAFE
Replicate Research Data Safely



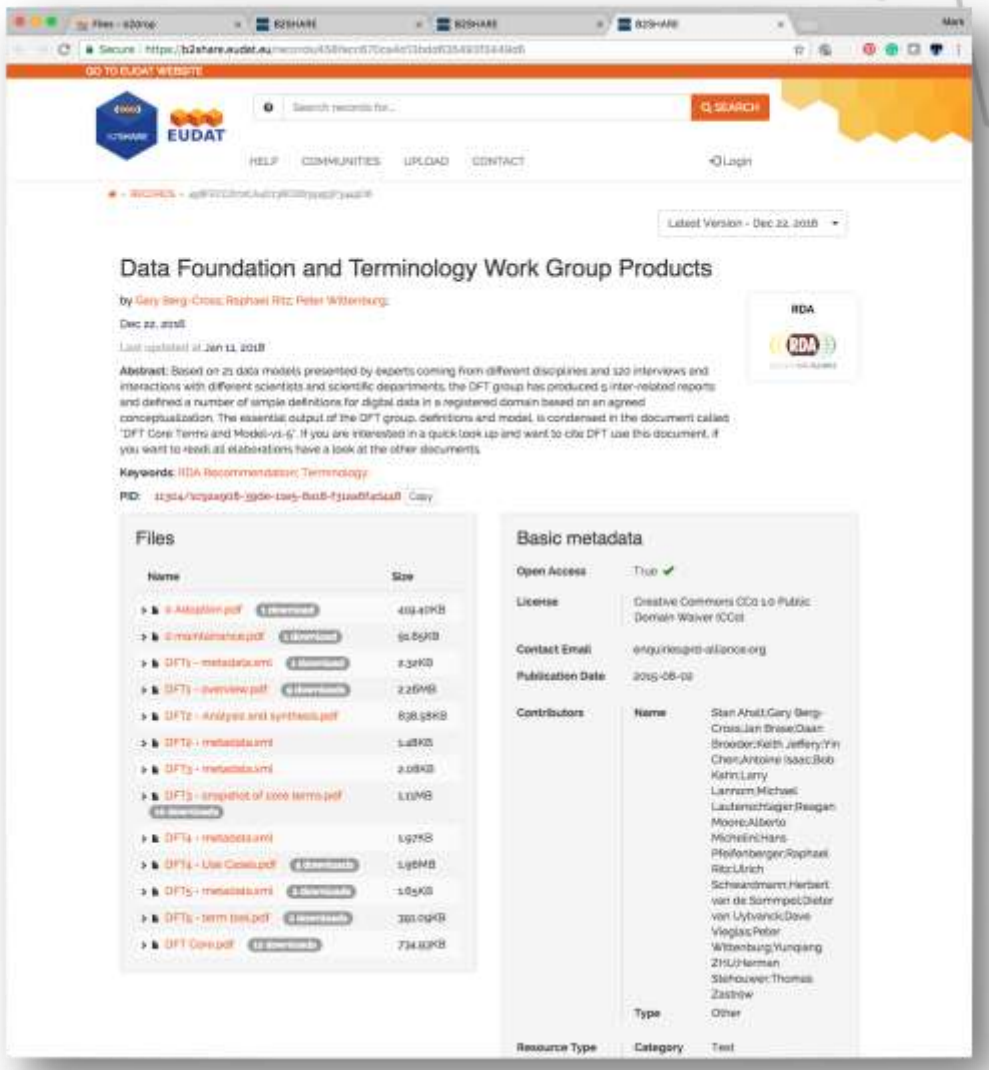
- **Who**
  - Community Data Managers
  - ‘Sophisticated’ Organizations
- **What**
  - Provide an abstraction layer which virtualizes large-scale data resources
  - Guard against data loss in long-term **archiving and preservation**
  - **Optimize access** for users from different **regions** and to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

SURF

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



https://b2share.eudat.eu/

Data replication & publication workflows

BSC

- Compute site
- B2SAFE endpoint

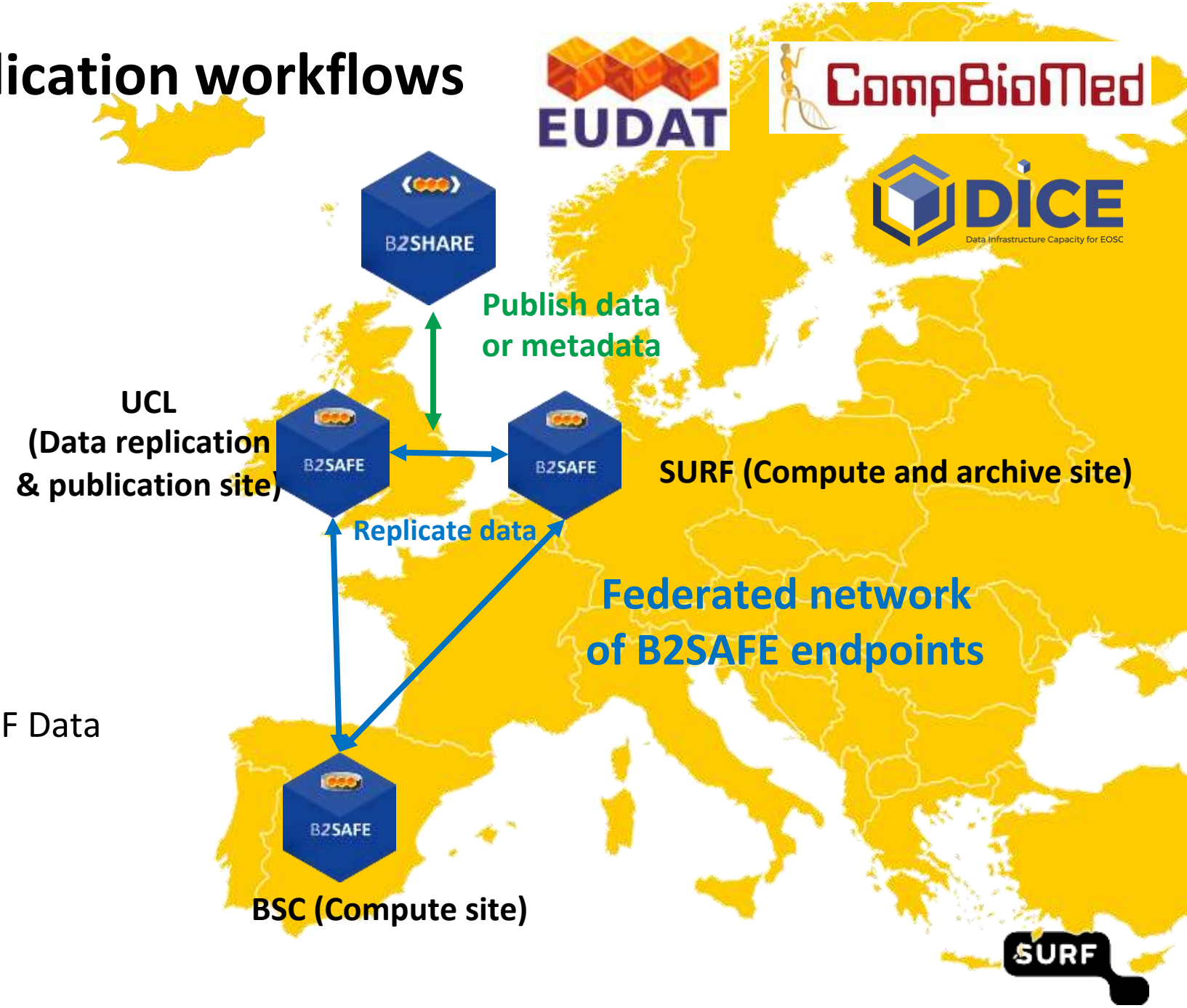
SURF

- Compute site
- B2SAFE endpoint
- Data Archiving site
- Possibility to publish data in SURF Data repository

UCL

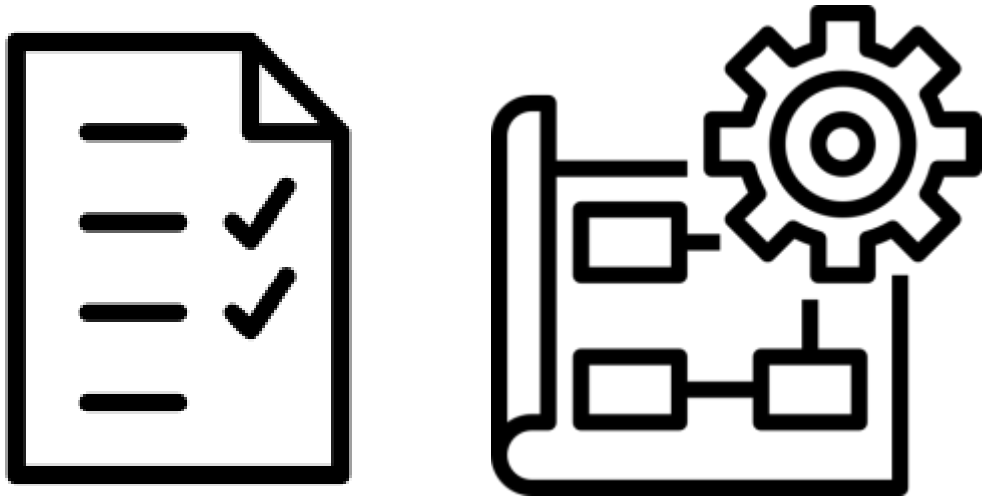
- Data publication site

  B2SAFE endpoint



Workplan and technical task descriptios

- We have made a workplan
- Started with deploying and configuration of services
- Technical support to deploy and using these services is provided through the CompBioMed and DICE collaboration



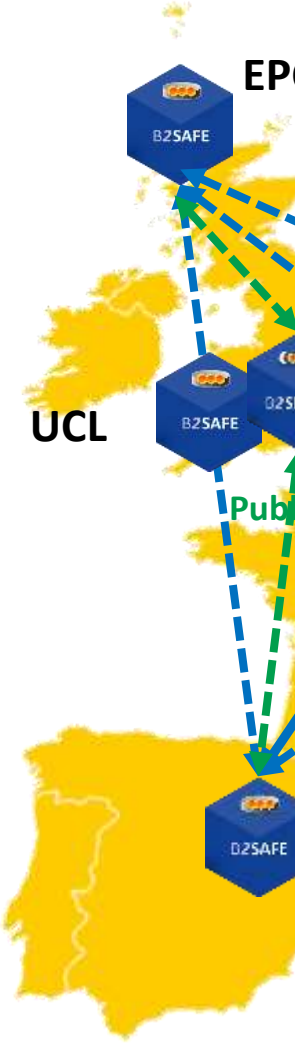
Detailed technical tasks

- BSC (Compute site)
- Deployment of B2SAFE tool
- Federation with other B2SAFE endpoints

- Allocation of storage in B2SAFE
- B2Handle or handle prefix (for making PIDs)
- SURF (Compute and archive site)
- Deployment of B2SAFE tool
- Federation with other B2SAFE endpoints
- Allocation of storage in B2SAFE and tape storage
- B2Handle or handle prefix (for making PIDs)
- Monitor integration of B2SAFE-B2SHARE
- UCL (Data publication site)
- Deployment of B2SAFE tool
- Federation with other B2SAFE endpoints
- Deployment of B2SHARE data repository
- B2Handle or handle prefix (for making PIDs)
- Integration B2SHARE-B2FIND

- Extend access to the platform to other HPC centers (e.g. LRZ, EPCC), research and medical centers in the community
- Safe data replication and data preservation
- Allocation of PIDs to replicated data
- Facilitate large data transfer
- Bring data close to compute
- Scale-up compute power
- B2SAFE-B2SHARE integration

3.4)



Metadata schema for CompBioMed community (addressed in CompBioMed Task

Thank you!