CompBioMed Use Case

EOSC Marketplace Ask me Anything webinar

Narges Zarrabi, SURF – The Netherlands

EOSC Future with C-SCALE DICE Data Infrastructure Capacity for EOSC EGI-ACE OpenAIRE NEXUS Reliance

The EOSC Future, C-SCALE, DICE, EGI-ACE, OpenAIRE-Nexus and Reliance projects are funded by the European Union Horizon Programme calls INFRAEOSC-03-2020 and INFRAEOSC-07-2020.

COMPBIOMED USECASE

ASK ME ANYTHING – SESSION 5: HPC - 7TH JUNE 2022

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

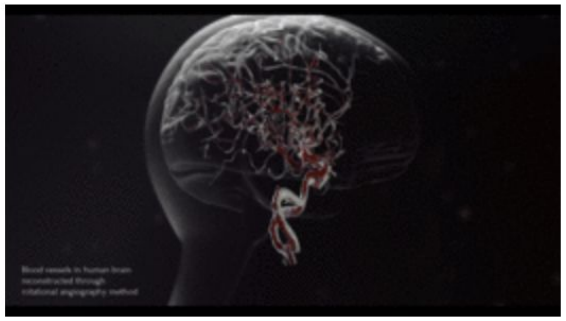
Training

Knowledge Exchange

SURF

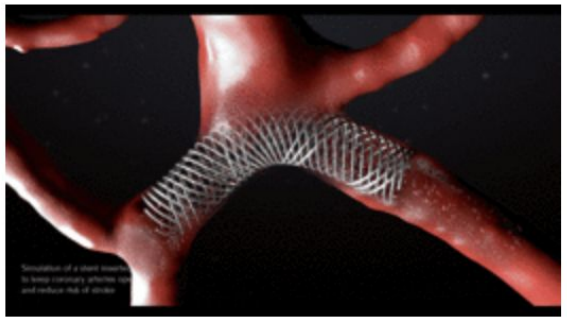
Computational BioMedicine

Academic Users



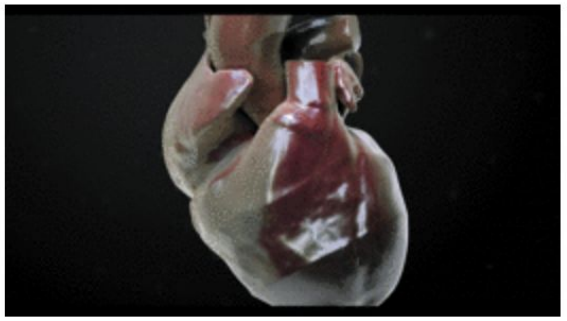
In this section you will find links relevant to Academic Users including user case studies, and information from our Academic Partners.

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

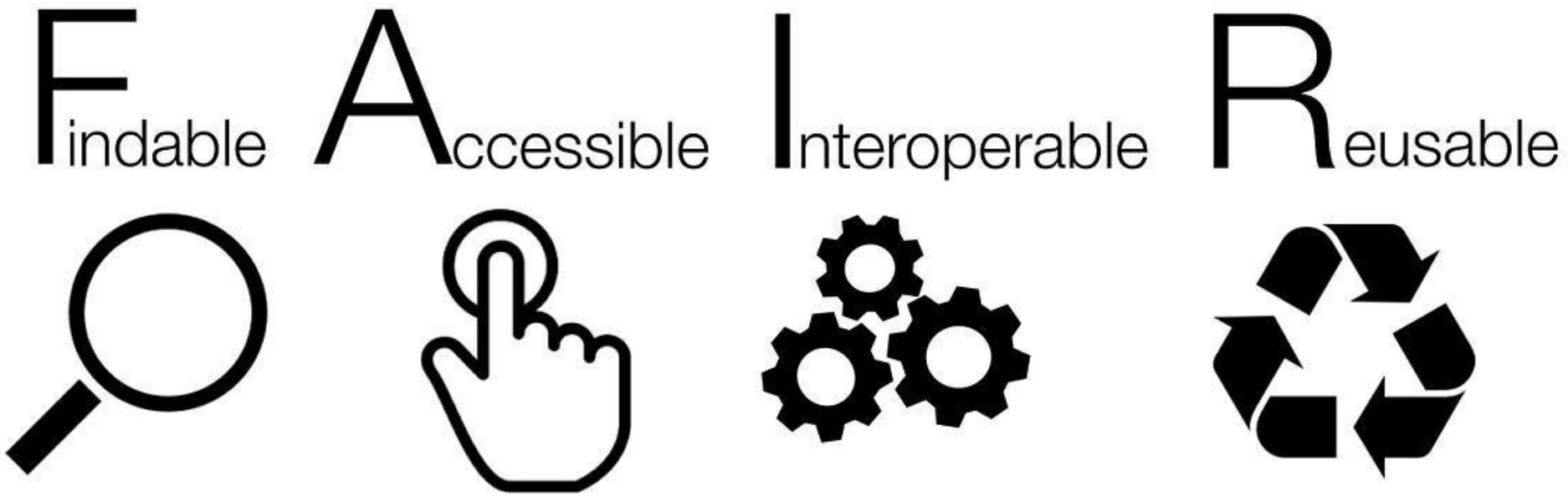
For those from the general public and media who are interested in our project and what we are planning follow this link and the relevant links on the page.

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*

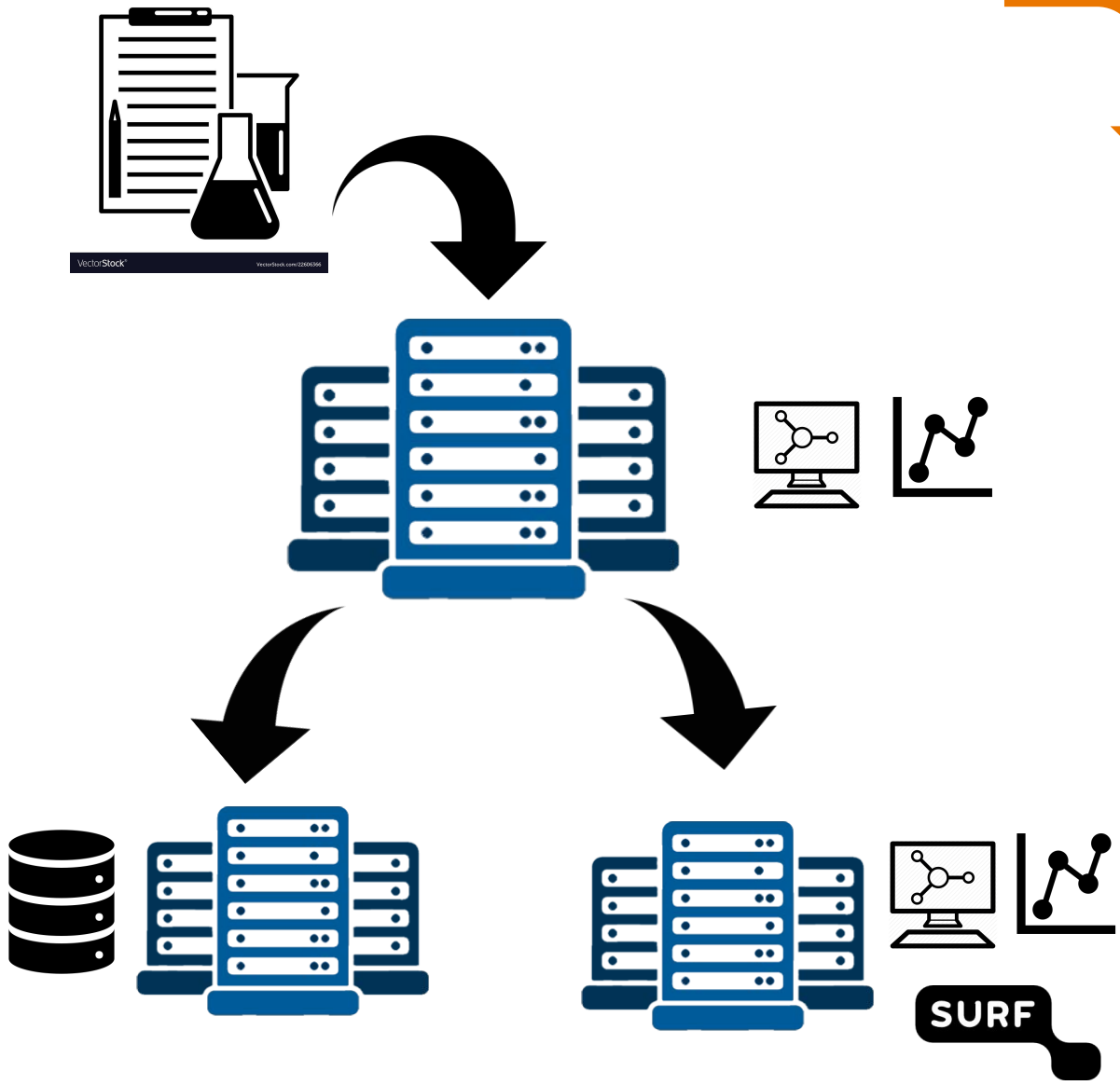
Main Challenge to make data FAIR

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



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



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

EUDADT services used in 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

Data & HPC Federation in CompBioMed

BSC

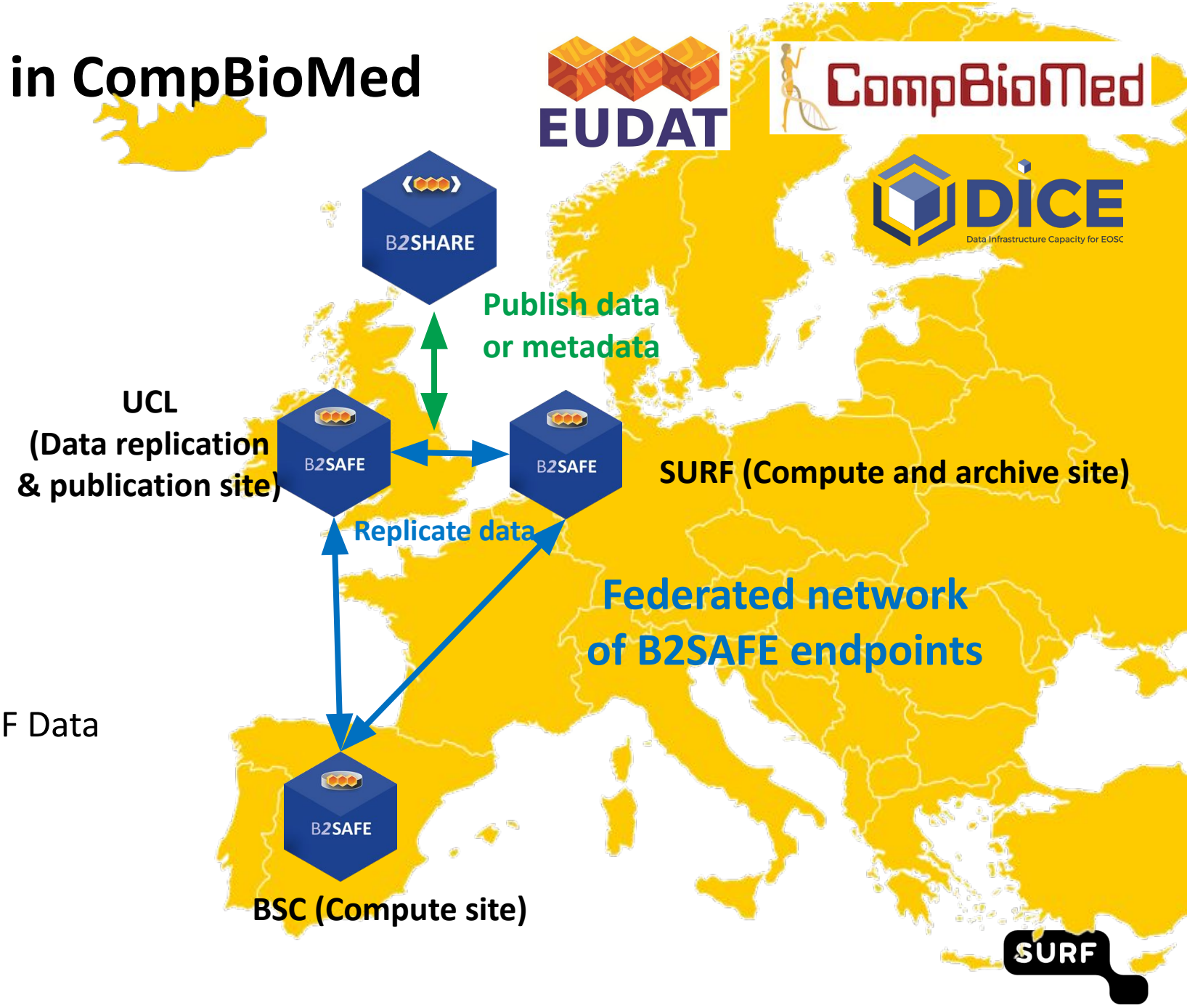
- Compute site
- B2SAFE endpoint

SURF

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

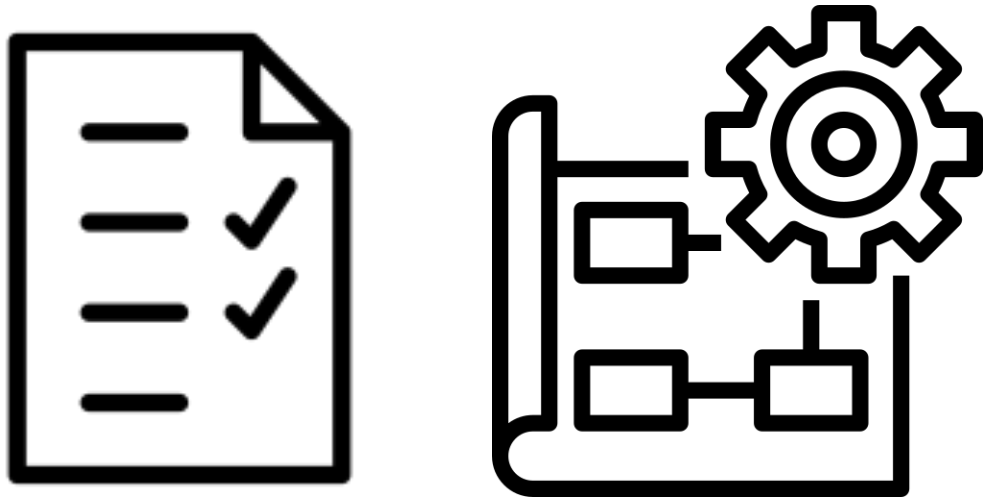
UCL

- 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

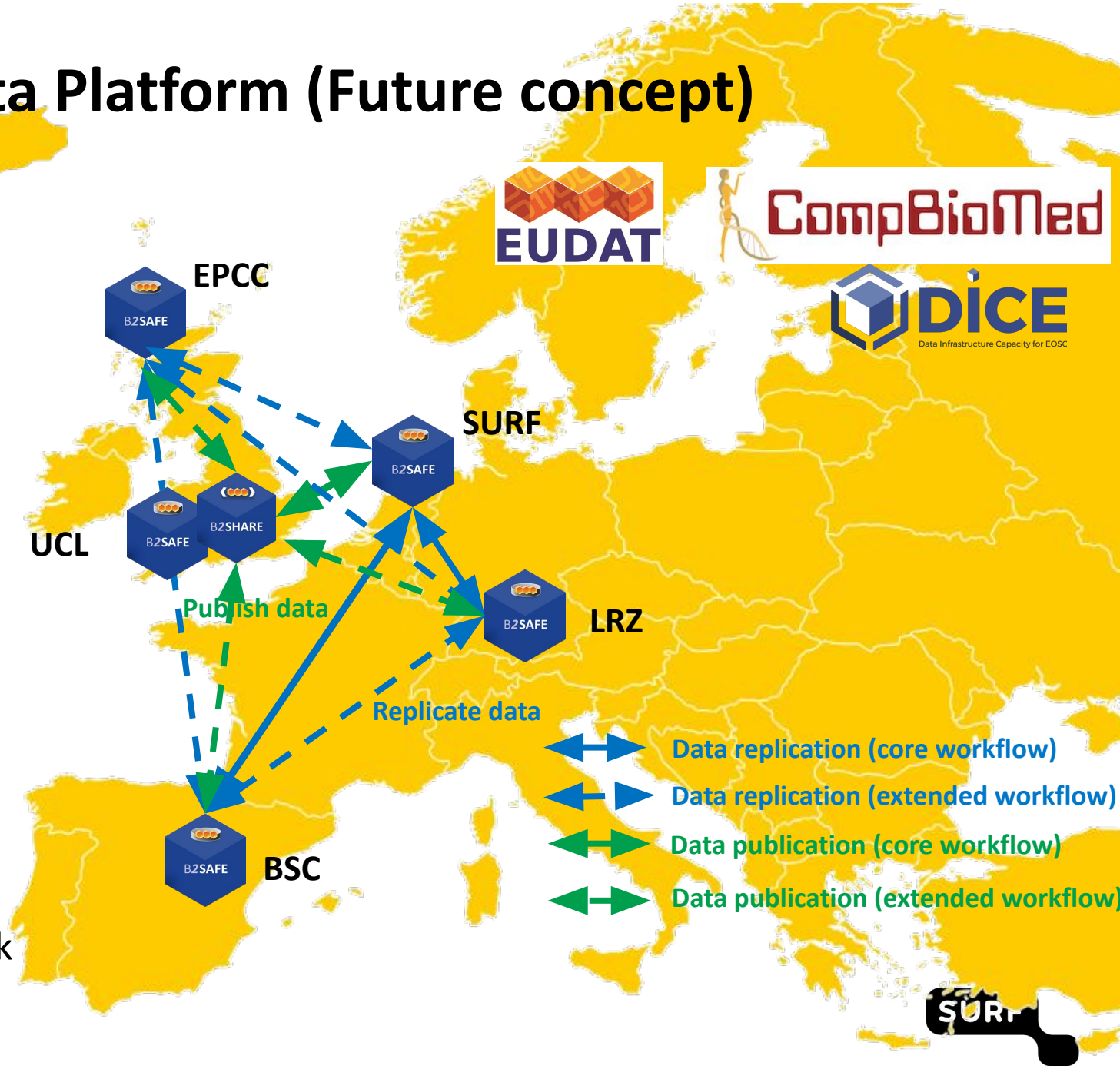


Detailed technical tasks

- 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

CompBioMed Federated Data Platform (Future concept)

- 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
- Metadata schema for CompBioMed community (addressed in CompBioMed Task 3.4)



Thank you!