EUDAT Services

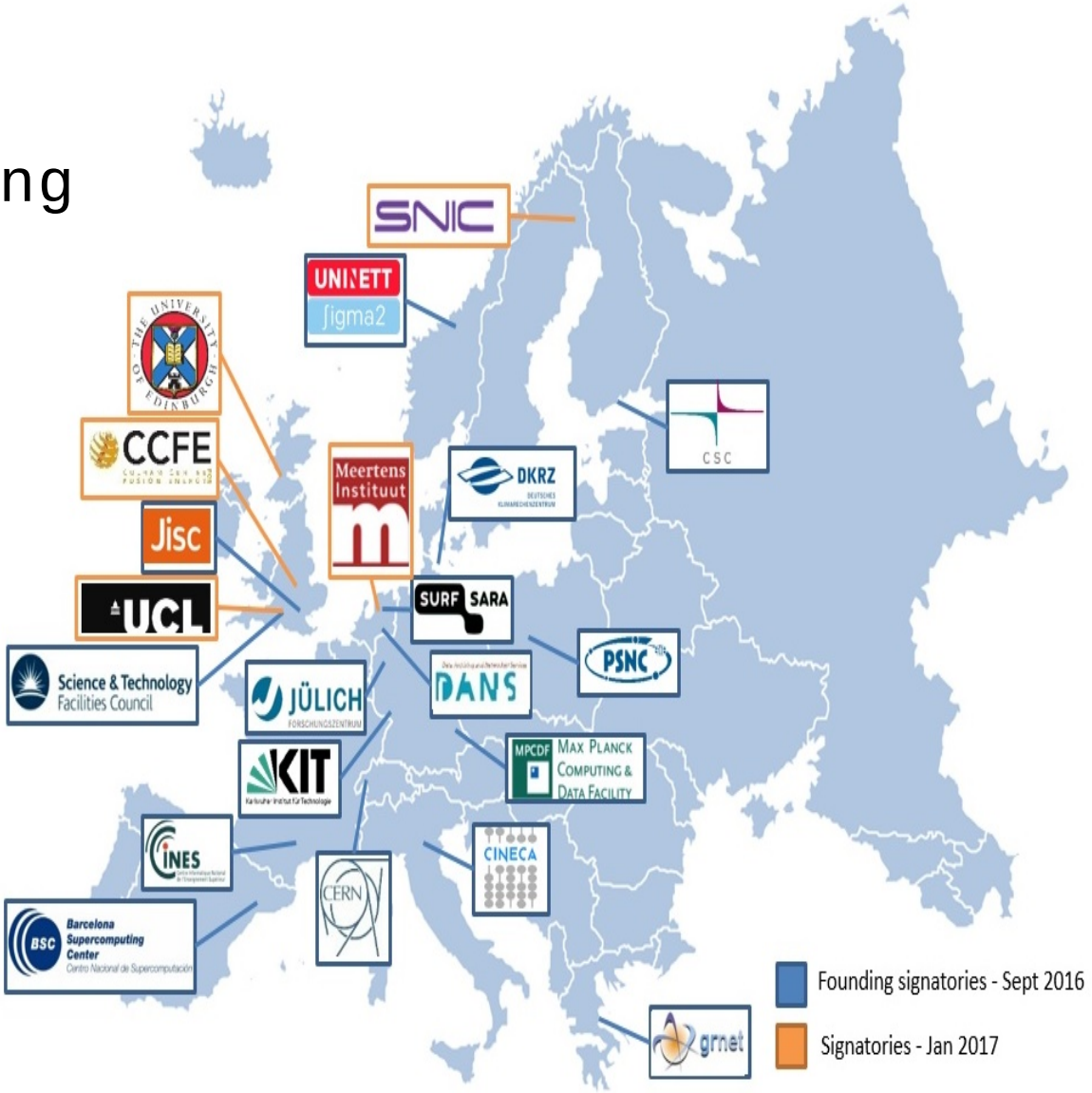
1st SeaDataCloud training workshop,
June 20 – 27, 2018, Oostende, Belgium

Sri Harsha Vathsavayi
CSC-IT Center for Science
Finland

EUDAT receives funding from the European Union's Horizon 2020 programme - DG CONNECT e-Infrastructures. Contract No. 654065

EUDAT Collaborative Data Infrastructure

A consortium of high performance computing (HPC)/ data centres, libraries, scientific communities, data scientists



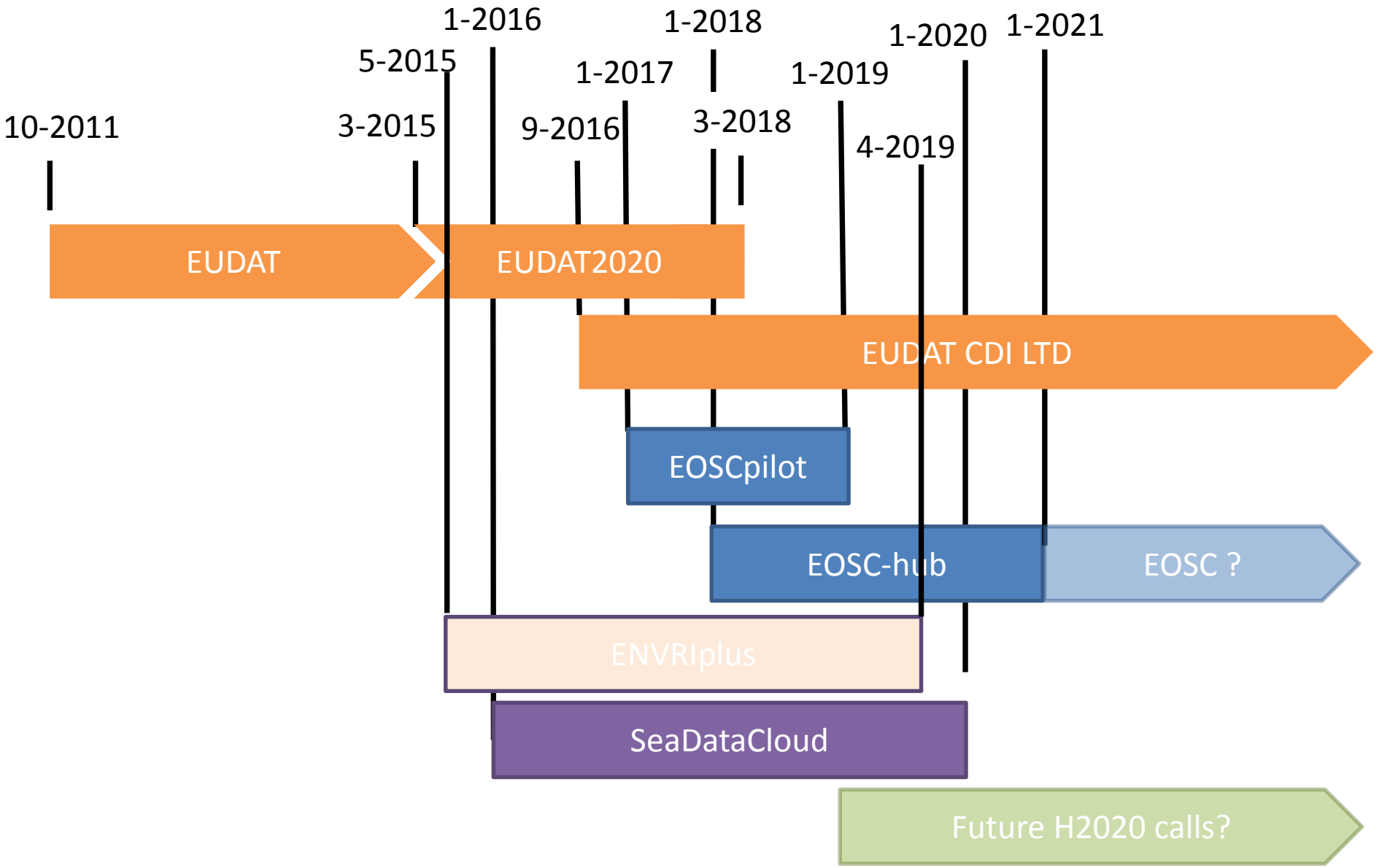
History of the EUDAT Collaborative Data Infrastructure

- EUDAT (1) project ended a long time ago
- EUDAT H2020 project ended February 2018
- EUDAT Collaborative Data Infrastructure (EUDAT Ltd.)

History of the EUDAT Collaborative Data Infrastructure

- Common Services for heterogeneous communities
  - Science data rates are exploding and will likely become continue to do so
  - Building bespoke services for new communities is not cost effective
- Initial Set of Services developed as result of community needs
  - Beyond the original ‘core’ communities
  - New services and specific community issues highlighted

Collaborative Data Infrastructure at a glance



EUDAT Core communities and Data Pilots


Direct simulation data of turbulent flows


High resolution simulations for Precision Cosmology


Tokamak data mirror for JET and MAST data


Unified Access to EISCAT radar data


Research data repository for students' own results


Data preservation and standardization in computational fluid dynamics

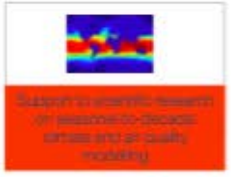

European Network for Earth System Modelling


European Long Term Ecological Research Network


Data storage and preservation of high-resolution climate experiments


European Plate Observing System


Distributed Research Infrastructure for Hydro-Meteorology


Support to scientific research on seasonal-to-decadal climate and air quality modelling


Dutch Atmospheric Data Distribution Service


Integrated Carbon Observation System

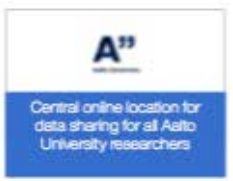

Public access to fine-grained city air quality data from roving sensors


Virtual Humans


Data Infrastructure for Chemical Safety


Information and Data Management Repository Platform for nanoscience in Europe


EUDAT as a long-term repository for the University of Porto


Central online location for data sharing for all Aalto University researchers


Enriching Europeana Newspapers


Cloud-like services to improve the preservation of digital cultural heritage


Storing, Cataloguing, Relating, and Exposing OCR Objects from the Open Philology Project


Common Language Resources and Technology Infrastructure


Data infrastructure for biodiversity and ecosystem community


Supporting data management for structural biologists

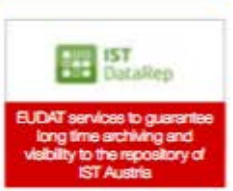

An EUDAT-based FAIR Data Approach for Data Interoperability


European Long Term Ecological Research Network


Linked Data service pilot

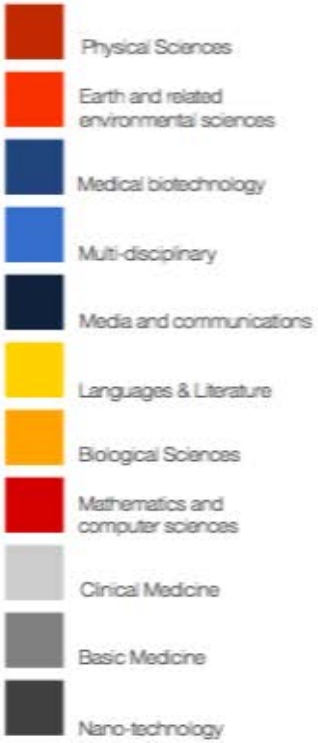

Long-term preservation of herbarium specimen images


A distributed infrastructure for life-science information


EUDAT services to guarantee long time archiving and visibility to the repository of IST Austria

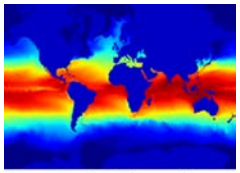

The use of the EUDAT repository to store clinical trials in a secure and compliant way


International Neuroinformatics Coordinating Facility



EUDAT & Environmental RIs

EUDAT ENV Data Pilots



Support to scientific research on seasonal-to-decadal climate and air quality modelling

Institutions and partners

For more information visit - https://eudat.eu/use-cases

Collaborative Data Infrastructure Services Suite

B2ACCESS
Identity & Authorisation

B2DROP
Synch and Share Research Data

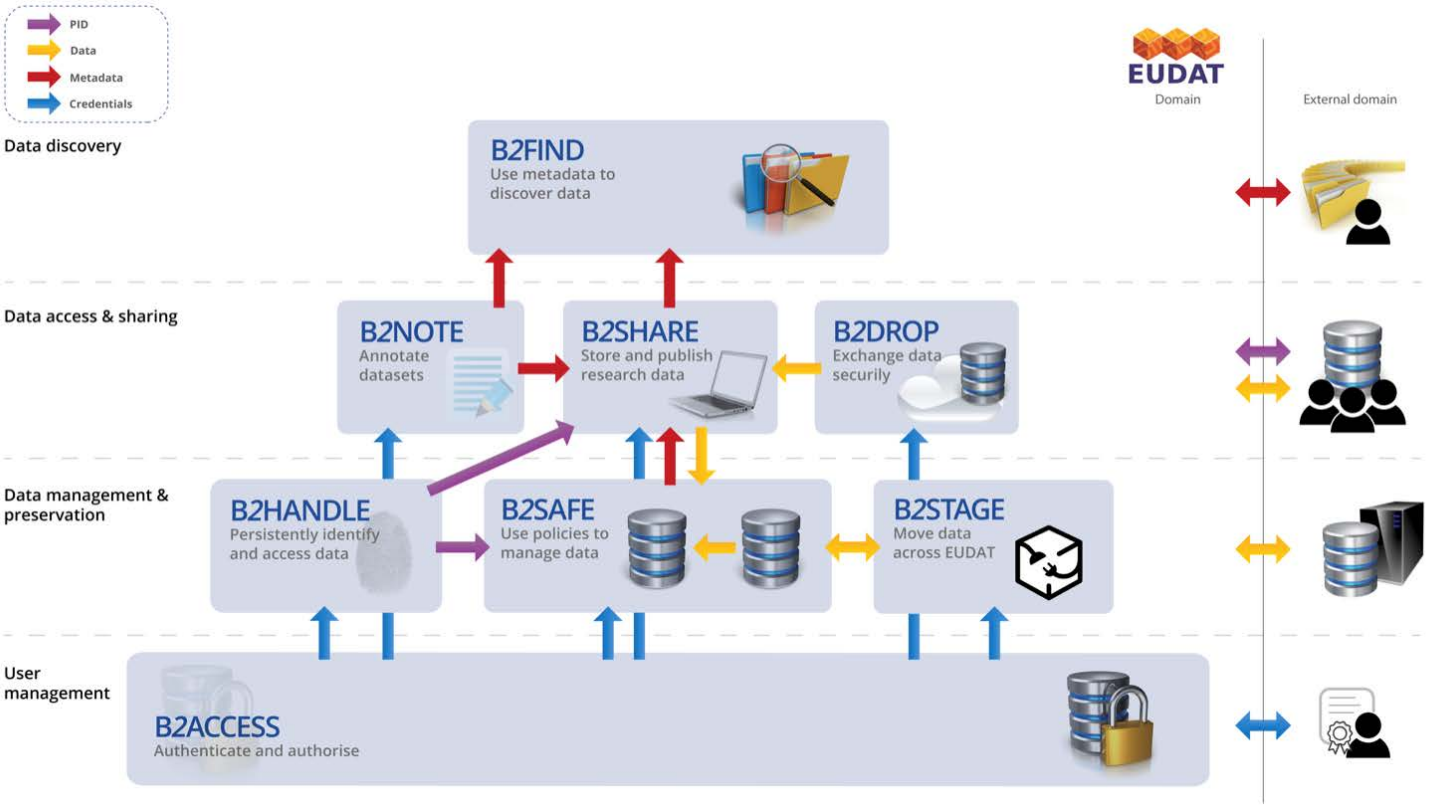
B2FIND
Find Research Data

B2SHARE
Store and Publish Research Data

B2SAFE
Replicate Research Data Safely

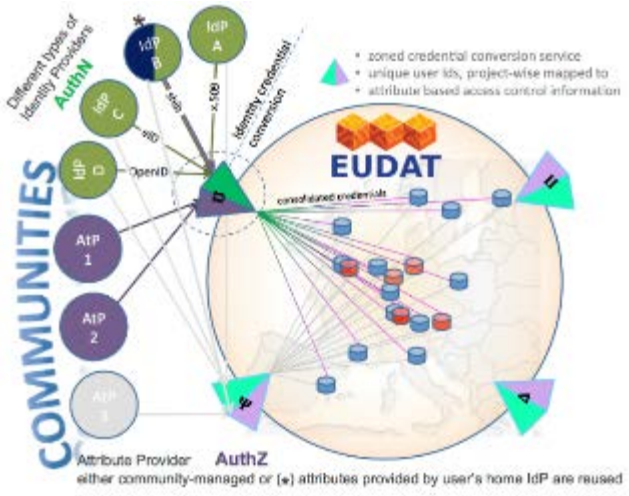
B2STAGE
Get Data to Computation

B2NOTE



http://www.eudat.eu/services

- **Who**
  - Anyone wanting to use the B2 Services
- **What**
  - Complies with **community ownerships** and **access rights**, basis of trust
  - Credential **conversion approach** (e.g. SAML, OpenID, X.509, Username/password)
  - Identity provider for **citizen scientists**
- **Why**
  - Use your own ID in federated environment

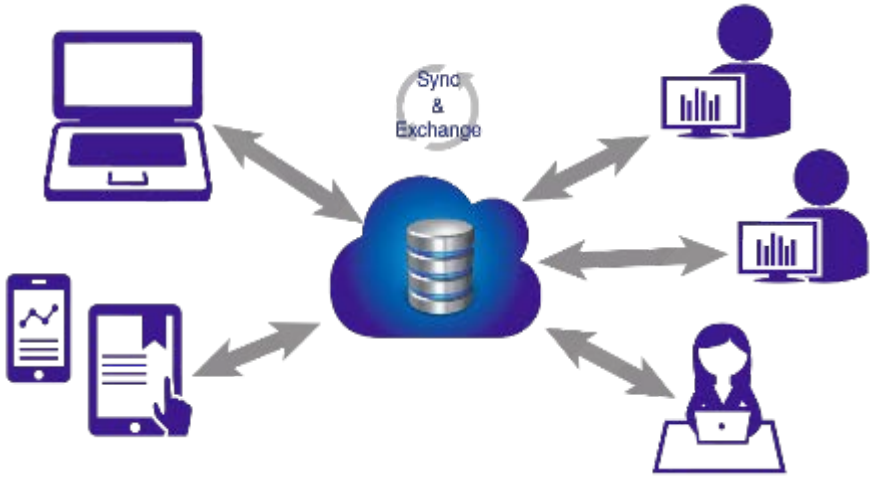


- **Stable/Production release of B2ACCESS, current version 1.9.6**
- **Add support for personal certificates to support eduGAIN, Social Identities (Facebook, Google, Microsoft, Gitlab), ORCID and local accounts**
- **IdP support for SAML, OpenID, OAuth2, X.509**
- **SP support for SAML, OIDC, OAuth2, X.509**
- **Integrated services: B2SHARE, B2SAFE, B2STAGE, B2DROP, B2NOTE, GEF, SPMT, DPMT, Confluence wiki, Gitlab,**
- **Pilot IdP integration with ELIXIR, PRACE, EGI**
- **Pilot integration with RCAuth CA**
- **EOSCpilot joint proposal for the Life Sciences with EGI and GEANT**
- **AARC pilot integration with GEANT eduTEAMS**

EUDAT

B2DROP
Sync and Exchange Research Data

- **Who**
  - Citizens Scientists and small teams
- **What**
  - **Store and exchange** data
  - **Synchronize** multiple versions
  - Ensure **automatic desktop** synchronization
- **Why**
  - Ease of Use
  - Trusted European Service



- **3 Major releases, ownCloud v9 and Nextcloud v11, v12**
- **Integration with B2SHARE via B2SHARE bridge app**
  - **2 Releases of the B2SHARE bridge app**
  - **To publish data stored in B2DROP to B2SHARE**
  - **multi file upload and selection of community domain**
- **Integration with B2ACCESS**

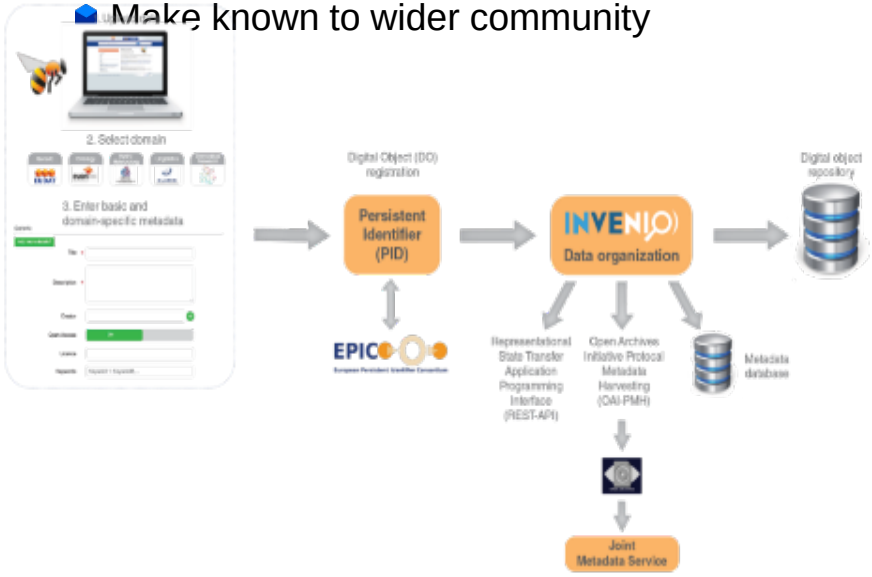
- **Who**
  - Anyone
- **What**
  - **Find** collections of scientific data quickly and easily, irrespective of their origin, discipline or community
  - Get quick **overviews** of available data
  - Browse through collections using **standardized facets**
- **Why**
  - Unique collection
  - Ease of Searching

- **Total 12 releases, 1 major release, current release 2.3.2**
- **Release B2FIND metadata schema, based on DataCite v3.0 and OpenAIRE guidelines**
- **Filter by Time (CKAN extension)**
- **Optimized workflow for harvesting and metadata ingest**
- **Extend harvesting methods to JSON-API and CSW**
- **19 (+6) Repositories harvested + 6 enabling + 12 planned**
- **Pilot integration with B2NOTE**

EUDAT

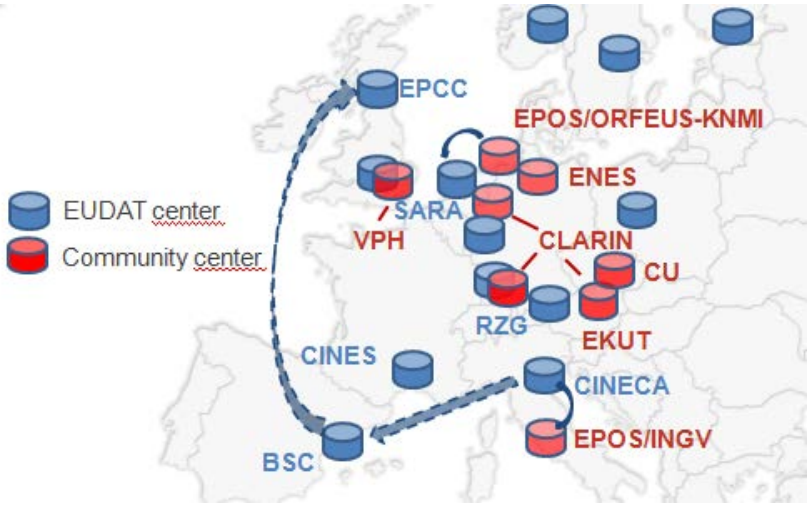
B2SHARE
Store and Share Research Data

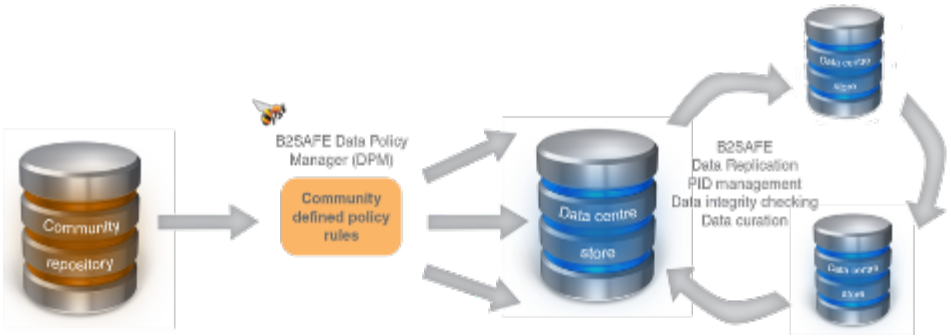
- **Who**
  - Small to Medium Teams
- **What**
  - **Store** data (incl. software) and add domain meta data
  - **Share** registered research data worldwide
  - **Preserve** (small-scale) research data for long-term
- **Why**
  - Register Data for Publications
  - Make known to wider community



- **Total 27 releases, 1 Major release B2SHARE v2, 13 RC for B2SHARE v2**
- **Current release v2.1.0**
- **B2SHARE v2 based on Invenio v3**
- **Support for DOI's and PIDs on object level**
- **Support for object checksums and verification**
- **Support for object statistics**
- **Support for Handle v8 HTTP API and B2HANDLE library**
- **Improved installation and upgrade procedures via Docker**
- **Flexible metadata model via JSON schema's for community domains**
- **Authorization to community domains**
- **Record lifecycle and versioning**
- **Integration with B2ACCESS, B2DROP and B2NOTE**

- **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
  - Bring data **closer to powerful computers**
- **Why**
  - Performance
  - Replication between trusted sites
  - Data Preservation



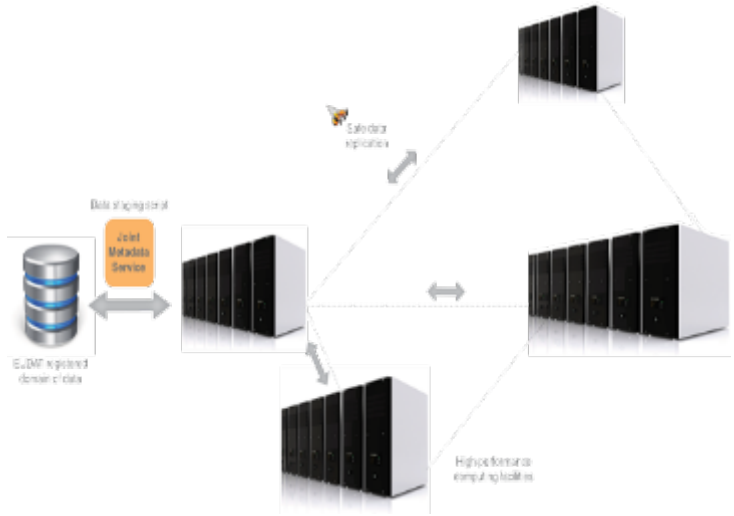


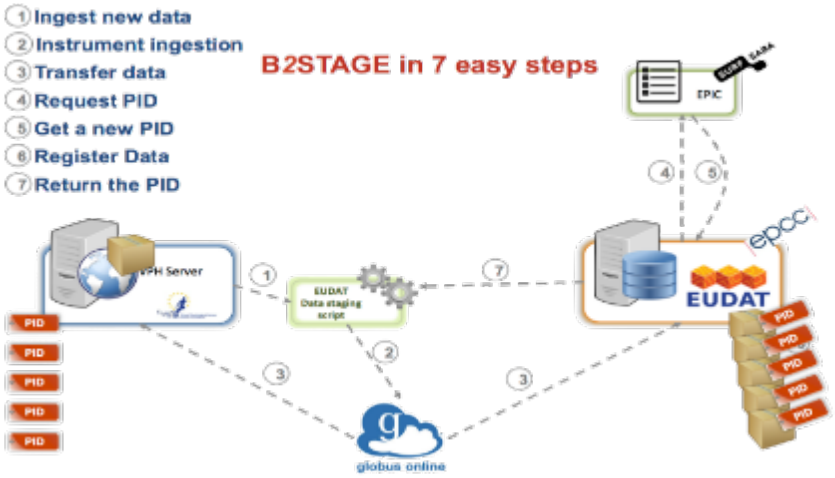
- Support **metadata**
- Optimize and **extend** policies to support **data curation** and **provenance**
- Support **authorization** on basis of **community access rules**
- Integration with other EUDAT services

EUDAT

B2STAGE
Get Data to Computation

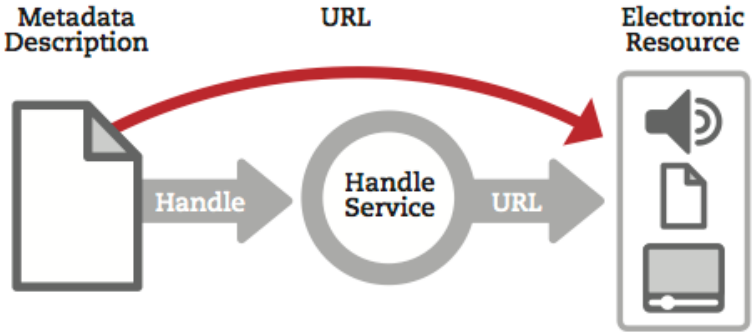
- **Who**
  - Users and Communities with Significant Computational Needs
- **What**
  - **Transfer** large data collections from EUDAT storages to **external HPC facilities for processing**
  - **Copy large data sets, ingesting them** onto EUDAT storage resources
- **Why**
  - Integration/Collaboration with PRACE & EGI
  - Simplify Data Transfer

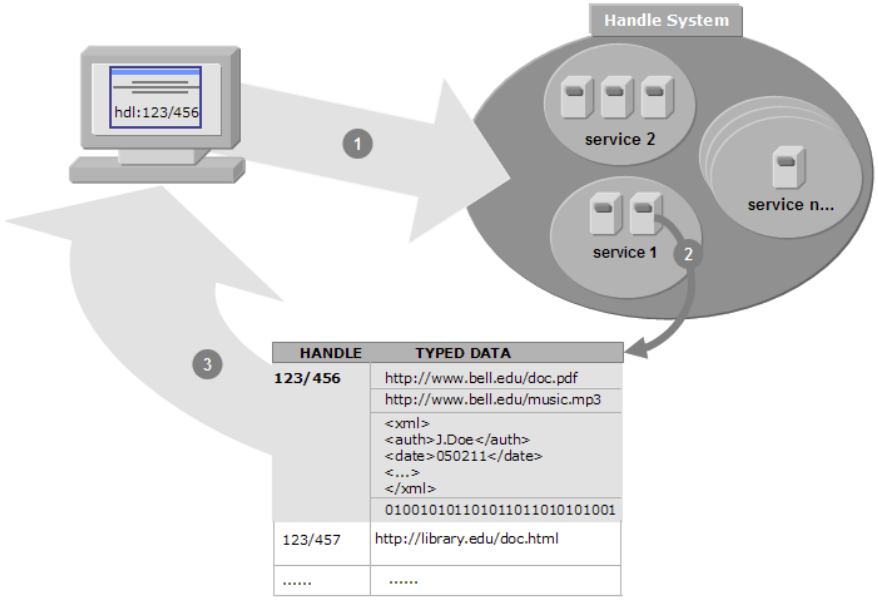




- Further develop **HTTP** to a **mature interface** and extend functionality to **metadata**
- Extend EUDAT **client API** library to other **B2 services** (e.g. B2SHARE, B2FIND, PID)
- Further integration with B2ACCESS

- **Who**
  - Groups or Communities who want to make their data citable
- What
  - Follows **policies** to register data and make it **long term refer**- and **citable**
  - Reliability through mutual **PID mirroring**
  - Provides **abstraction layer** between a globally **unique persistent identifier** and **physical location** of data objects
  - M**achine readable** via **HTTP RESTful API**
- **Why**
  - Simple integration
  - Technology Agnostic



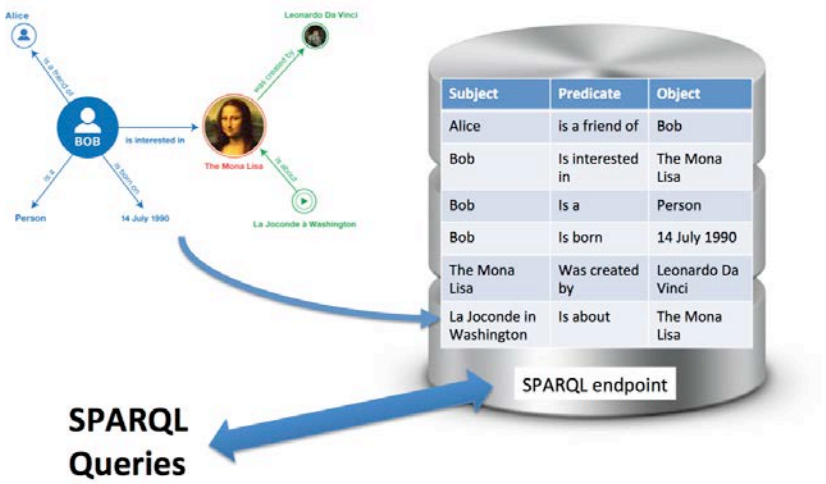


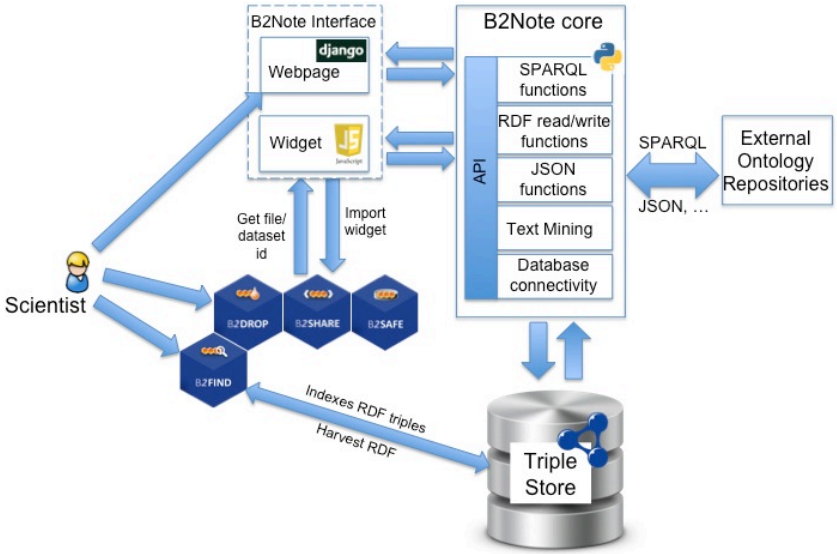
Development plan

- Develop the policies for the B2HANDLE service (e.g. PID namespace mngmt)
- Migrate service from Handle v7 to v8
- Define PID Information Types for data, metadata, collection records
- Integrate with Data Type Registry service
- Consolidate B2HANDLE API library with EUDAT API library

Features

- Creation **RDF triples**
- Harvests information from **ontology** repositories
- Supports **semi-automatic annotation** using text mining
- Supports **manual data annotation**
- Easy to use **user interface**
- **Integrates** with the different **B2 services**





Development plan

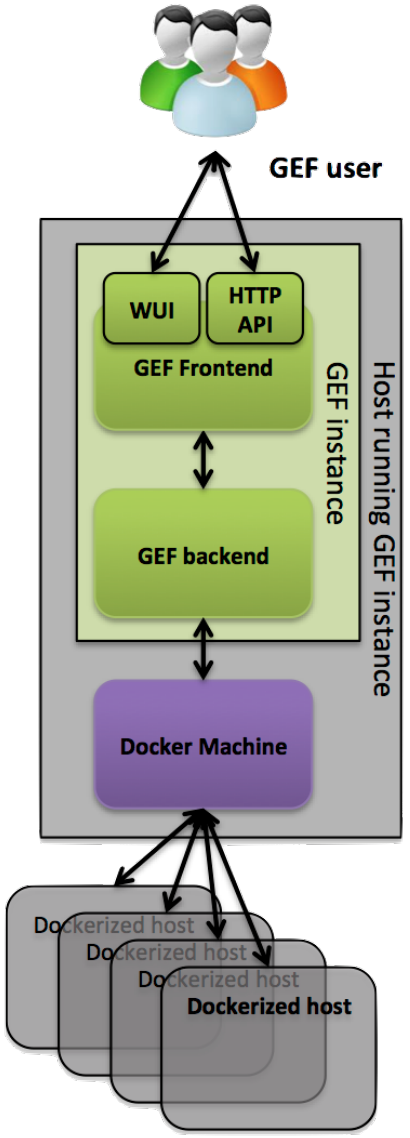
- UI for manual annotation, initial focus on annotation of **metadata**
- Setup and test **Triple store**
- Develop **harvesting chain** for **ontologies**
- **Integration** with **B2SHARE** and **B2FIND**
- Assess the use of **Graph** technologies

Some New Services in Development

Generic Execution Framework

DATA SUBSCRIPTION

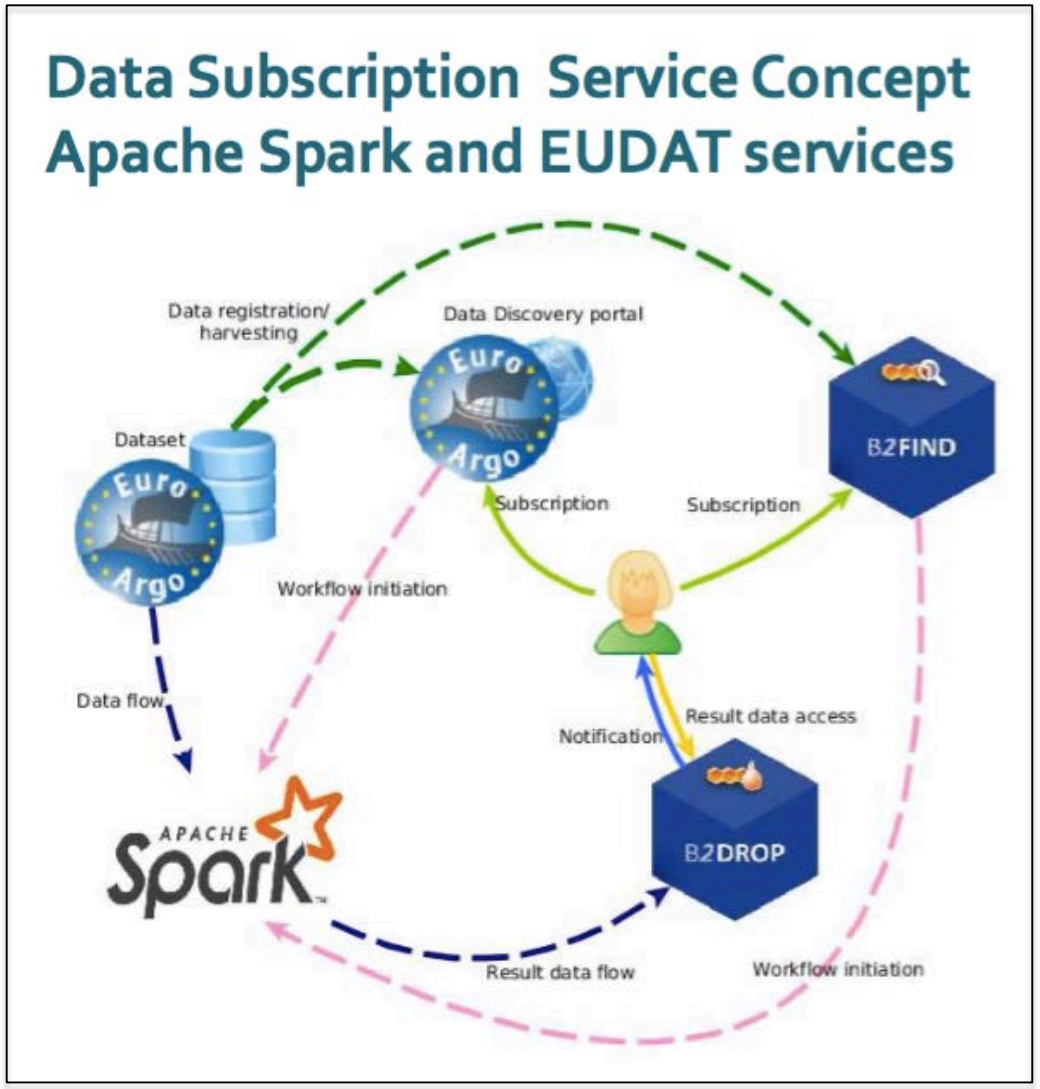
- Minimize data transfers
- Move tools, not data
- Enable data processing close to the data
- User provide community specific execution engines (e.g. docker containers)
- Stimulate reproducible results, reuse of execution templates
- Integration with CDI data services, containers to access B2SHARE, B2DROP, B2SAFE
- Integration with B2ACCESS



EUDAT

Use Case

- Aims at a **generic** solution
- Data storage providers and subscribing applications are **inside** and **outside** the **EUDAT CDI domain**
- **Data storage providers** and **applications** representing individual users **subscribe** to data through a well-defined interface
- Subscriptions are activated by **matching notifications**

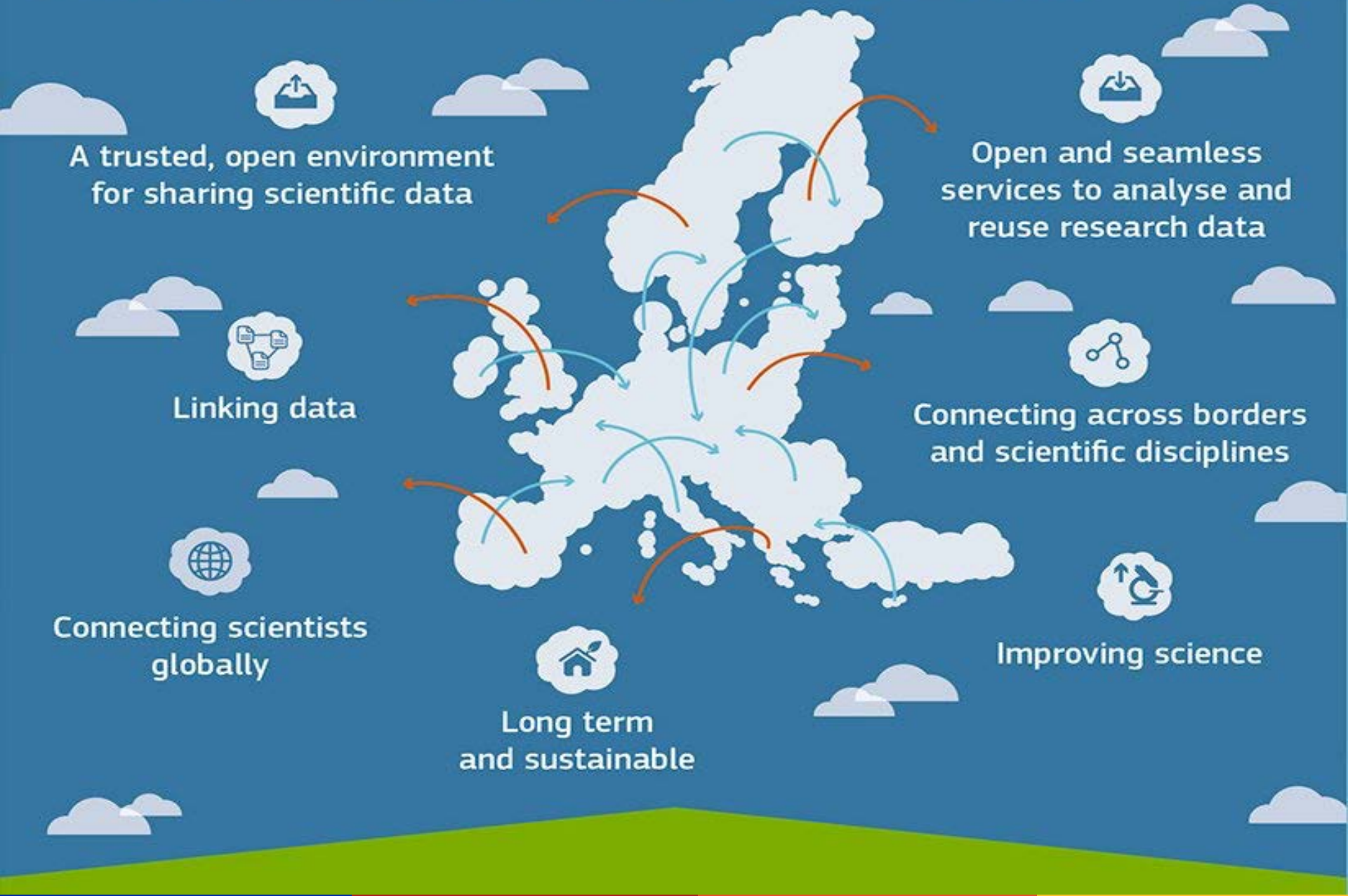


Development plan

- **Integration** with **a community data selection UI**
- **Develop** further the subscription matching algorithm
- **Integration** with **B2ACCESS**

EUROPEAN OPEN SCIENCE CLOUD

BRINGING TOGETHER CURRENT AND FUTURE DATA INFRASTRUCTURES



EOSC-hub

EOSC-hub mobilises providers from 20 major digital infrastructures, EUDAT CDI**, EGI* and INDIGO-DataCloud jointly offering services, software and data for advanced data-driven research and innovation.

** CDI – Collaborative Data Infrastructure

* EGI is not an acronym (any more)

EOSC-hub Mission

The project will create **EOSC Hub:**

a federated integration and management system for EOSC

To establish a unified, open set of production services realized via cross-Infrastructure integration, procurement, delivery, innovation and training, involving multiple providers and benefitting research collaborations of pan-European and international relevance.

EOSC-hub and OpenAIRE

EOSC-hub is collaborating with OpenAIRE to promote OpenScience in Europe

For details about OpenAIRE, visit https://www.openaire.eu/

Service Providers

e-Infra

Generic services

Humanities

- Language and literature (CLARIN)
- Arts (DARIAH)

Engineering

- Environmental engineering (sea vessels, LNEC)
- Civil Engineering (Disaster Mitigation)

Medical and Health Sciences

- Biological Sciences (ELIXIR)
- Structural biology (WeNMR)

Natural sciences

Physical Sciences

- Astronomy (LOFAR)
- Fusion (ITER)
- High Energy Physics (CMS and VIRGO)
- Space Science (EISCAT-3D)

Earth Science

- EO Pillar
- GEO
- Climate Research (ENES)
- Seismology (ORFEUS, EPOS)

Biological Sciences

- Marine and freshwater biology (IFREMER)
- Biodiversity conservation (LifeWatch)
- Ecology (ICOS)

Community services & pilots

Thematic service providers

- ✓ CMS, astronomy/astro-particle physics
- ✓ Climate change/ENES
- ✓ GEO/Global Earth Observation System of Systems
- ✓ Costal protection/LNEC
- ✓ Structural biology/WeNMR
- ✓ Earth Observation core data resources/ESA, Copernicus sentinel data, LANDSAT and other major EO data archives
- ✓ Biodiversity / Lifewatch

Competence Centres

- ✓ Marine research
- ✓ LOFAR
- ✓ EuroFusion/ITER
- ✓ EPOS/EIDA seismic data and computational seismology services
- ✓ EISCAT-3D
- ✓ ELIXIR/BBMRI/ECRIN
- ✓ DARIAH
- ✓ ICOS

EOSC-Hub Marine CC

- **Competence Center – Early Adopter**
- **Objective**: provide an open data analytics platform fully supported by e-infrastructures where scientific users will be able to select, filter, aggregate, synthesize large volume of reference marine observations.
- Specific objectives:
  1) aggregate distributed contents and data from multiple marine RI sources together;
  2) facilitate researchers interactive access and analysis of data;
  3) build a sustainable service for the years to come

Next Steps?

- Foster new pilots & joint activities (e.g. upcoming H2020 calls → "*INFRAEOSC calls: Stepping up the establishment of the EOSC and connecting ESFRI infrastructures through Cluster projects*" (?)
  - Leverage existing work with e-Infrastructures → we already have some building blocks!
- Facilitate the adoption & integration of RIs to and within the EOSC
- Future H2020 calls

Contact information:

Chris Ariyo Chris.Ariyo@csc.fi
Service Manager, Research Data Services, CSC – IT Center for Science, Finland

Damien Lecarpentier damien.lecarpentier@csc.fi
Project Director, Research Infrastructures, CSC – IT Center for Science, Finland
EUDAT Project Director

https://b2(service).eudat.eu/
http://www.eudat.eu/support-request