CSC

ICT Solutions for Brilliant Minds

The role of CSC in Open Science in Finland, OS/FAIR infrastructures and tools

Hanna Koivula

2024-04-15

Our special expertise includes for instance research infrastructures, interoperability, and digital transformation

Turnover in 2022

64 M€

One of the world's most eco-efficient datacenter in Kajaani

Non-profit state enterprise with special tasks owned by the state of Finland **70 %** and Finnish higher education institutions **30 %**

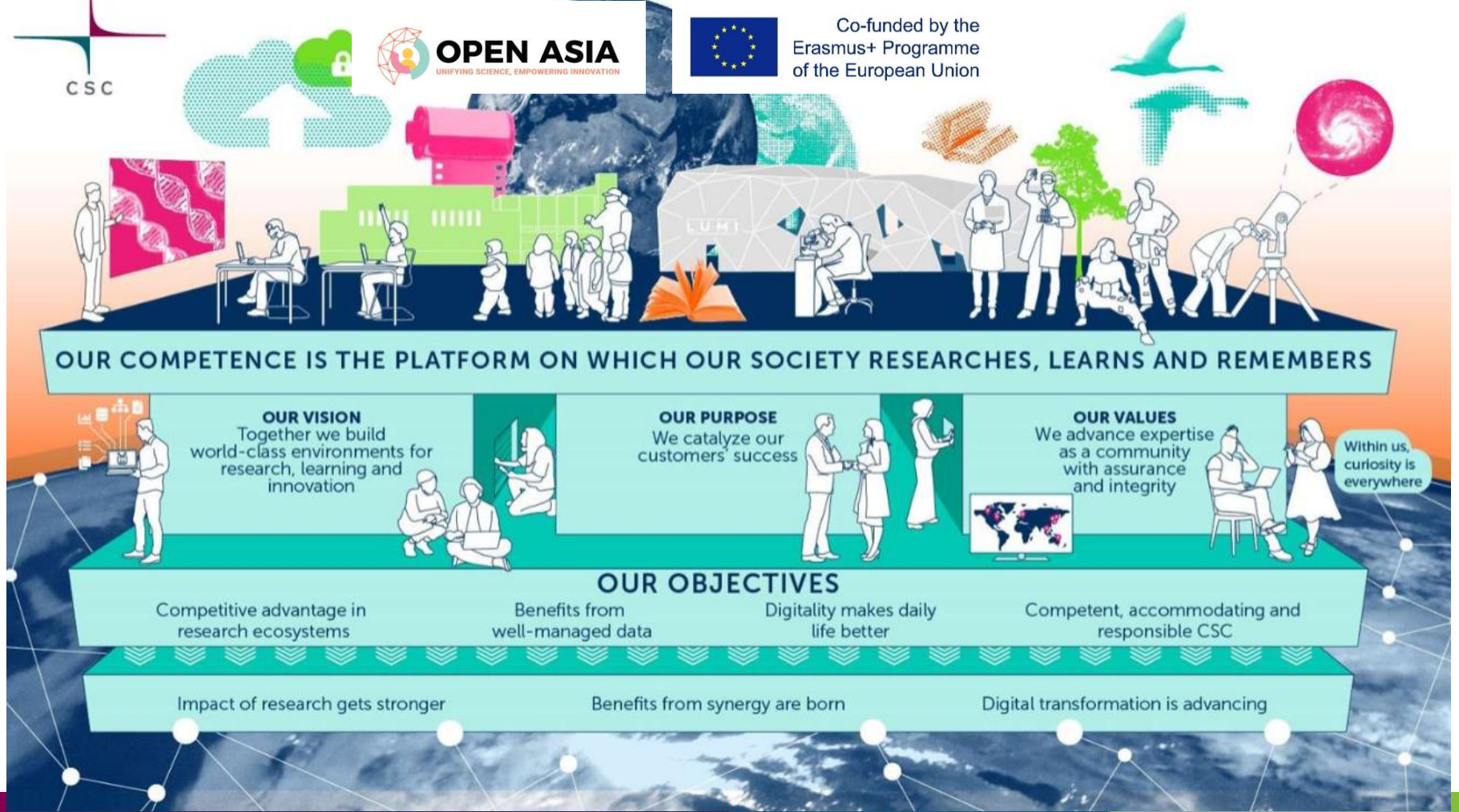
OPEN ASIA
UNIFYING SCIENCE, EMPOWERING INNOVATION

Co-funded by the Erasmus+ Programme of the European Union

Approx.

670

employees in 2024


CSC
OPEN ASIA
UNIFYING SCIENCE, EMPOWERING INNOVATION
Co-funded by the Erasmus+ Programme of the European Union
LUMI
OUR COMPETENCE IS THE PLATFORM ON WHICH OUR SOCIETY RESEARCHES, LEARNS AND REMEMBERS
OUR VISION
Together we build world-class environments for research, learning and innovation
OUR PURPOSE
We catalyze our customers' success
OUR VALUES
We advance expertise as a community with assurance and integrity
Within us, curiosity is everywhere
OUR OBJECTIVES
Competitive advantage in research ecosystems
Benefits from well-managed data
Digitality makes daily life better
Competent, accommodating and responsible CSC
Impact of research gets stronger
Benefits from synergy are born
Digital transformation is advancing

Open Science and FAIR Guiding Principles

Open Science is about incentives and actions (Data Governance), but more concrete guidance, and techincal solutions are needed for its execution (Data Management)

?

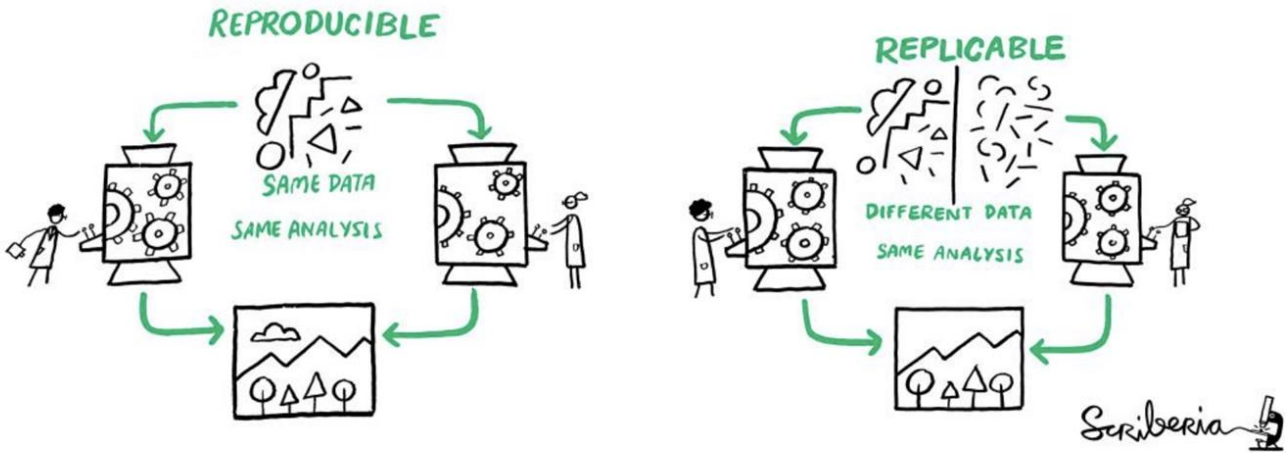
The **FAIR Guiding Principles** bring in more spesific definitions for **data quality, security, interoperability, transparency and machine actionability**

?

FAIR gives more concrete means to get the **benefits of the open data**

Reproducibility of science in digital age FAIR

- A fundamental principle of the original scientific method i.e. part of the so called "good scientific practice"
- A continuum: **reproducibility replicability re-usability**



***Source**: Reproducibility of scientific results in the EU – Scoping report*, Lusoli, W.(editor), Publications Office, 2020, **https://data.europa.eu/doi/10.2777/341654**

Understandable for Humans + Machine Actionable

F

FINDABLE

- Essential information described in sufficient detail (**metadata**)
- Description page and has a persistent identifier (**PID**)

A

ACCESSIBLE

- **Can be searched** on the Internet
- **Versioning and life cycle** are documented
- **Tombstone page** if data has been deleted

I

INTEROPERABLE

- Common, documented and **open file formats** are used
- **Data content and constraints** are also interoperable
- Structure and content use **standards and vocabularies**

R

RE-USABLE

- **Data quality is well documented** and understandable
- **Access rights** displayed and machine actionable

Wilkinson, M., Dumontier, M., Aalbersberg, I. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci Data* **3,** 160018 (2016). https://doi.org/10.1038/sdata.2016.18

OPEN ASIA
UNIFYING SCIENCE, EMPOWERING INNOVATION

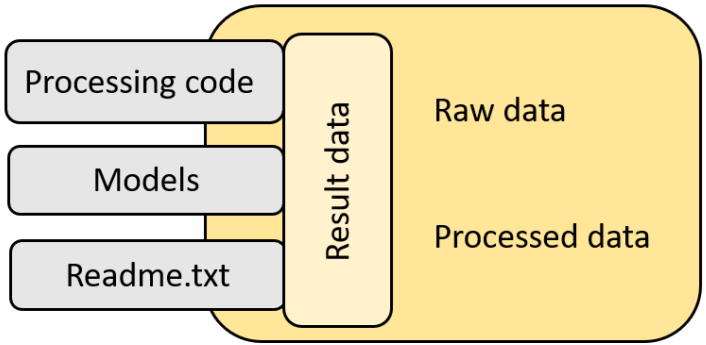
Co-funded by the Erasmus+ Programme of the European Union

FAIR principles apply to all research output(s)

Discovery metadata (open)

- Administrative
- Descriptive
- Structural

Data documentation documents & data (open, restricted, controlled)

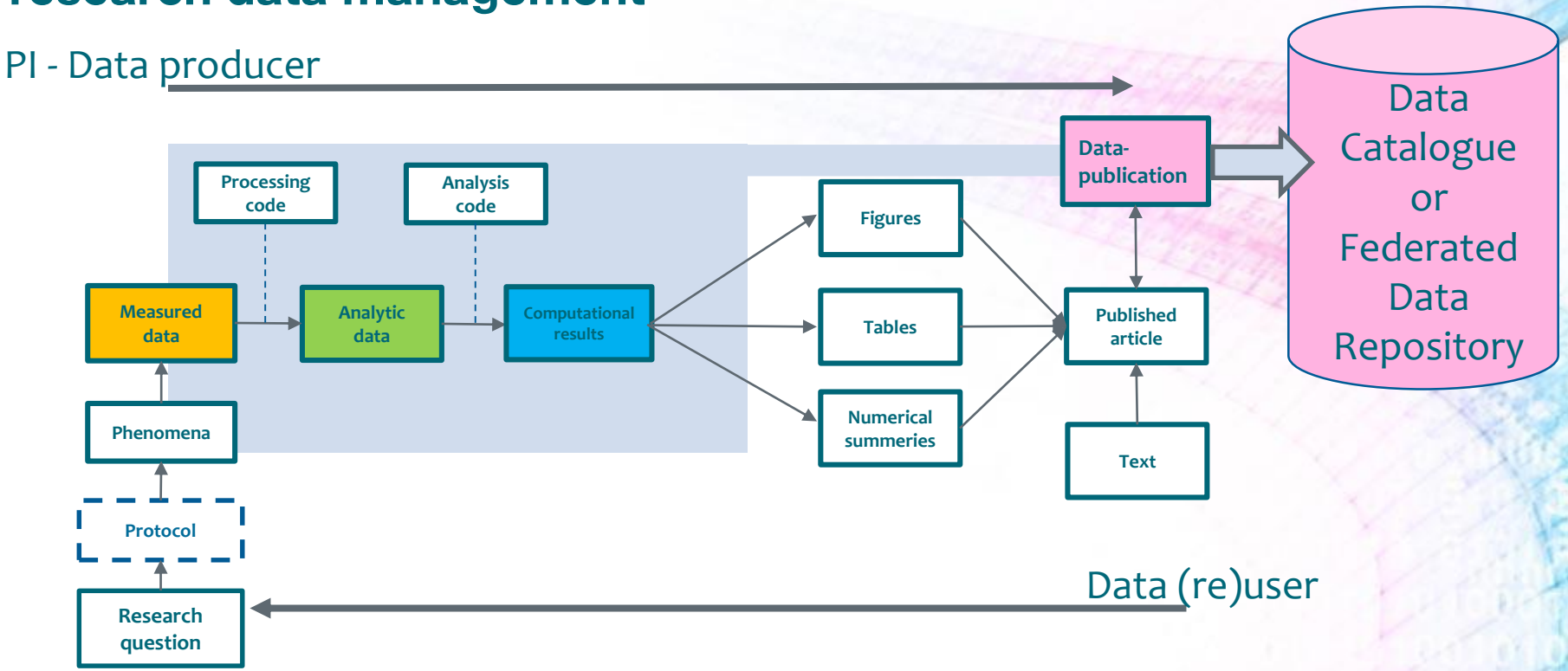


Public information → **Open license / Terms of use / Controlled access**

Data and can be open, restictedly available or controlled. Terms of use (for research) have been described in the administrative metadata and machine readable lisence.

FAIR:
”Data should be as open as possible and as restricted as necessary”

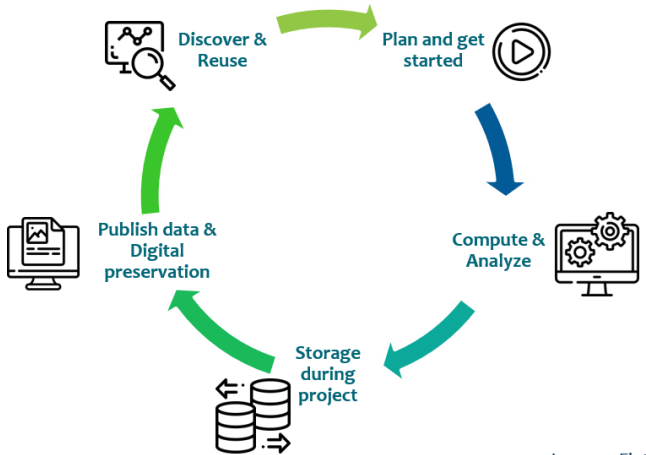
Enlarging the impact of research with systematic research data management



OPEN ASIA
UNIFYING SCIENCE, EMPOWERING INNOVATION
Co-funded by the Erasmus+ Programme of the European Union

FAIR by design - from collection to publishing and re-use

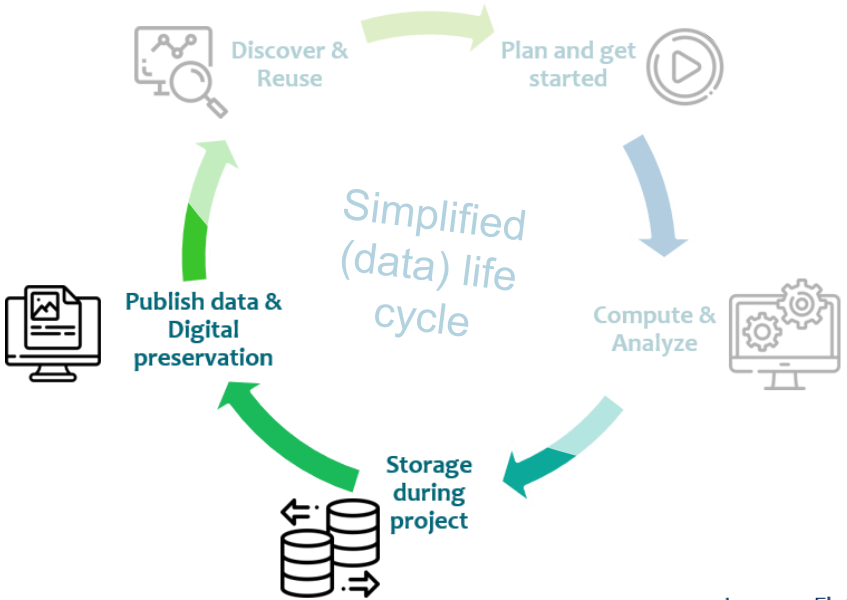
- **Data Management Plan (DMP)**
  - Describes tools and services needed during the entire data life-cycle
- **Metadata and documentation**
  - Data protection and GDPR
  - Agreements and licenses
  - Document workflows and automate as much as possible
- **Data**
  - Quality Assurance (and documentation)
  - Standards and vocabularies (if available)
  - Version control
- **Analysis**
  - Computational capacity, AI and modelling
- **Publishing**
  - Choose a publishing platform that is sustainable
  - Make data citable (with PIDs)
- **Storage solutions after completing research**
  - Federated data services
  - Long-term preservation



Images: Flaticon.com

What do we mean by storing, sharing and publishing data?

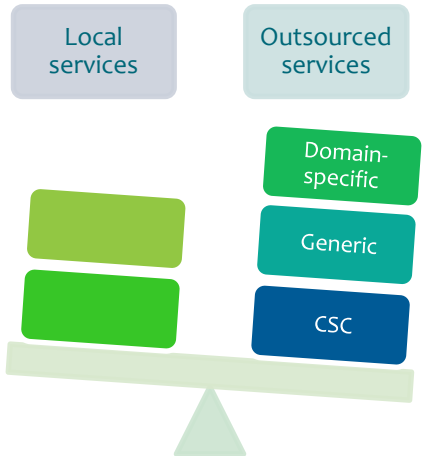
- **Storing:** storing data in a service for a period of time
- **Sharing:** sharing data through shared storage space or by using links
- **Publishing** data as dataset: Dataset metadata is available for others via a landing page, that has a persistent identifier
- **Digital preservation**: reliable preservation of digital information for several decades or even centuries



Images: Flaticon.com

Things to consider for the user / organisation when choosing services

- ✓Service description and promise
  - ✓Sustainability
  - ✓Data security
  - ✓Version control
  - ✓DOIs and other PIDs
- ✓Domain or data type specific requirements for (meta)data
  - ✓Is it suitable for personal and/or sensitive data?
  - ✓Domain specific Quality Assurance tools
  - ✓Where to find standards and vocabularies?
- ✓Amount of data
  - ✓Local or cloud services?
  - ✓Local server or HPC



OPEN ASIA
UNIFYING SCIENCE, EMPOWERING INNOVATION
Co-funded by the Erasmus+ Programme of the European Union

Organising the national research ecosystem – to ensure FAIR by design

- Collaboration and common understanding of user requirements – common architecture
- Building skills in national and /or scientific networks
- Establishing joint competence center and career paths for different roles
- Participating in international networks like RDA or CODATA
  - Research Data Alliance RDA - https://www.rd-alliance.org/
  - Committee on Data International Science Council CODATA https://codata.org/
- Setting targets and measuring
  - FAIR maturity models and frameworks help in setting targets
  - Sertificates eg. Core Trust Seal for data repositories

OPEN ASIA
UNIFYING SCIENCE, EMPOWERING INNOVATION
Co-funded by the Erasmus+ Programme of the European Union

OPEN ASIA
UNIFYING SCIENCE, EMPOWERING INNOVATION

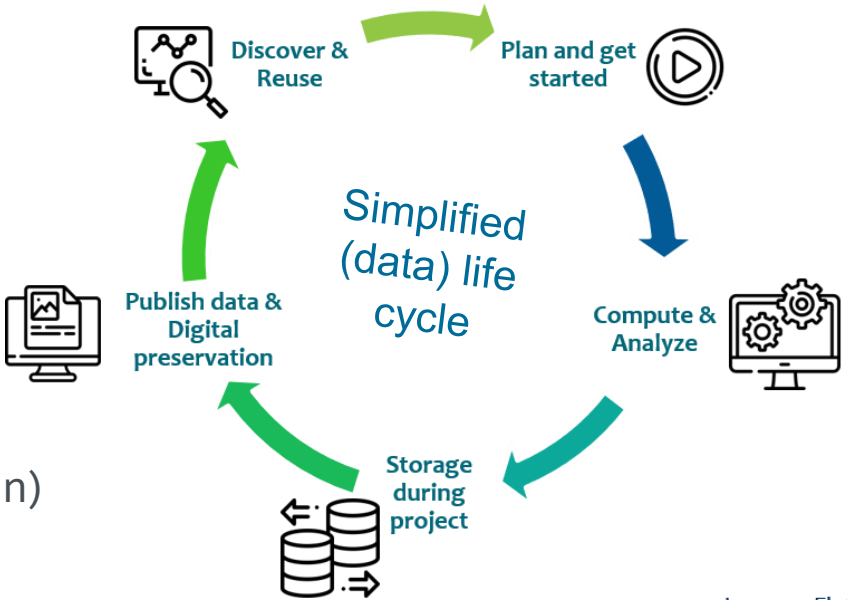
Co-funded by the Erasmus+ Programme of the European Union

CSC's research data management services

FAIR by design?

FAIR services cover the entire data life-cycle:

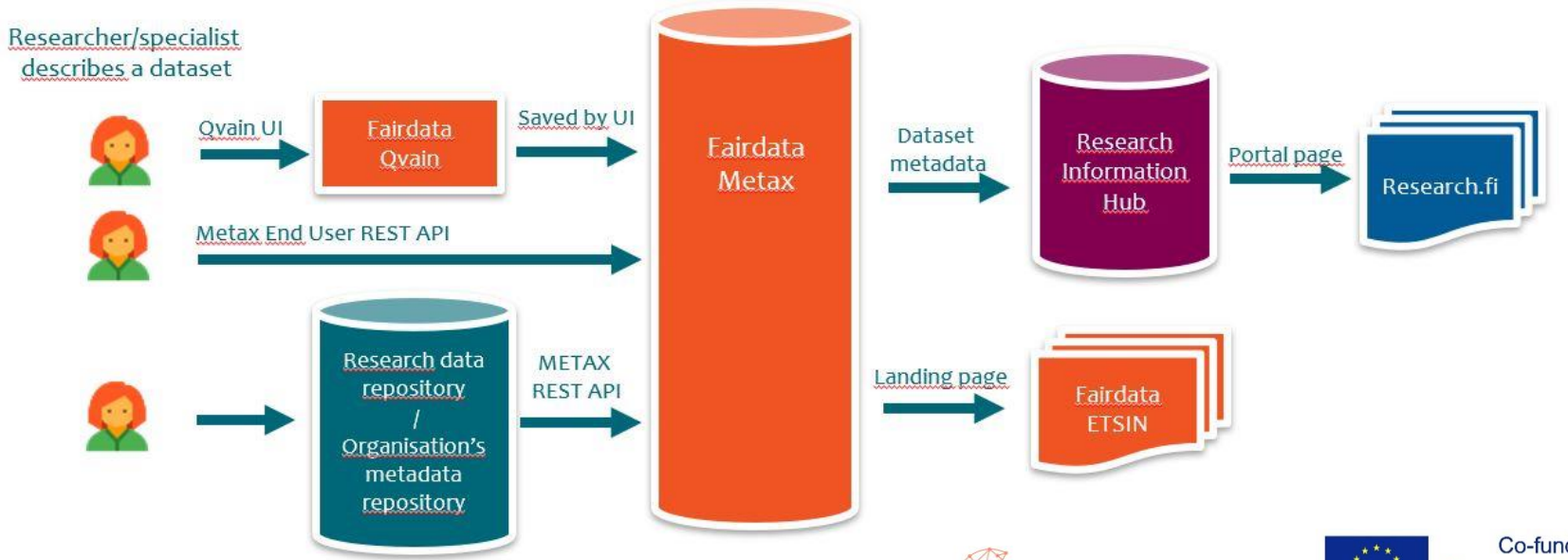
1. Research plan is the basis for DMP
2. Data collection, agreements and IPR
3. Compute/process and analyze
4. Storage during project
5. Publish data (+ Long Term Preservation)
6. Discover and reuse



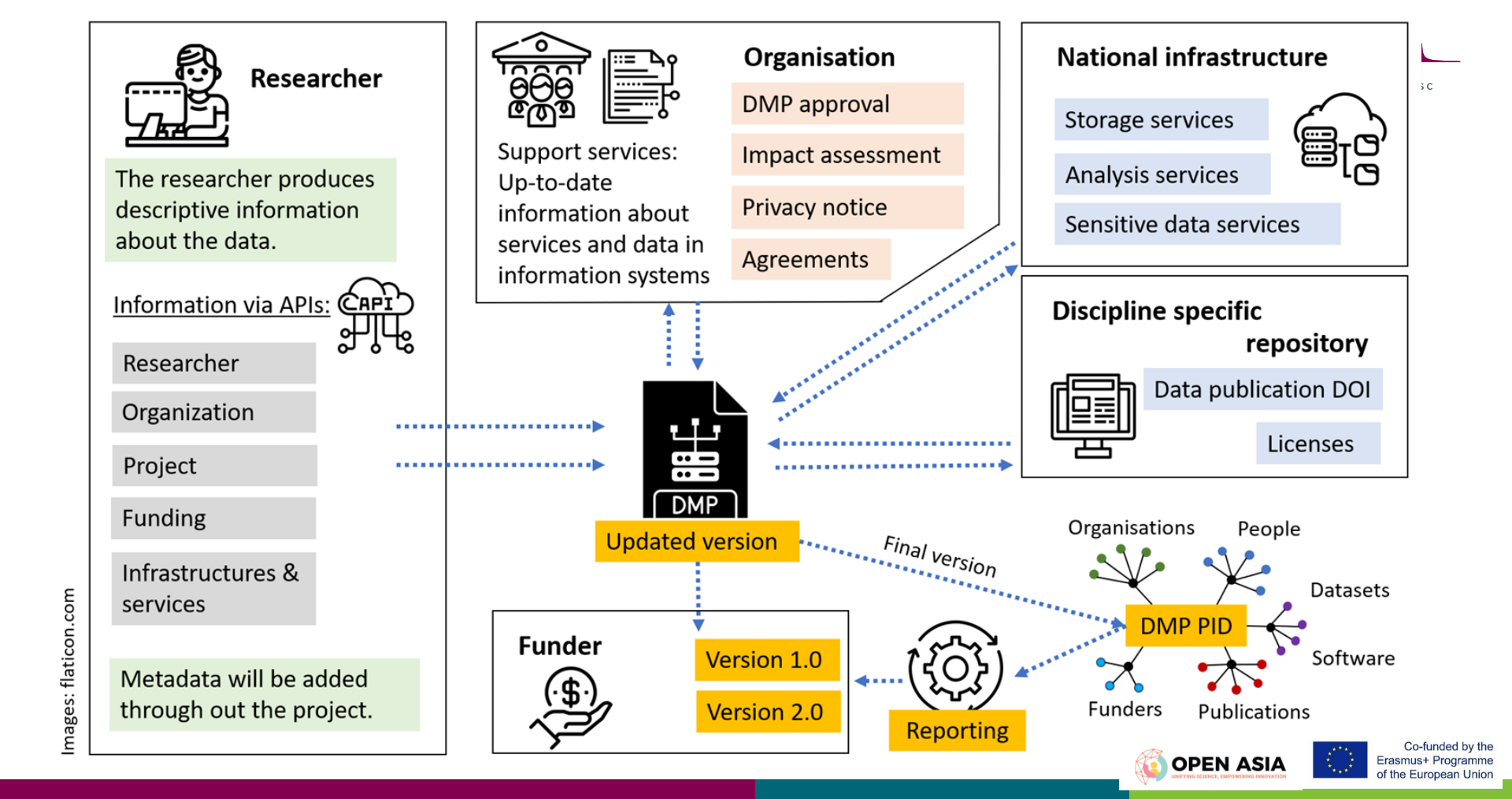
Images: Flaticon.com

CSC's services are bind together with METAX metadatabase

Metax doesn't have a graphic web user interface, but can be used over APIs.

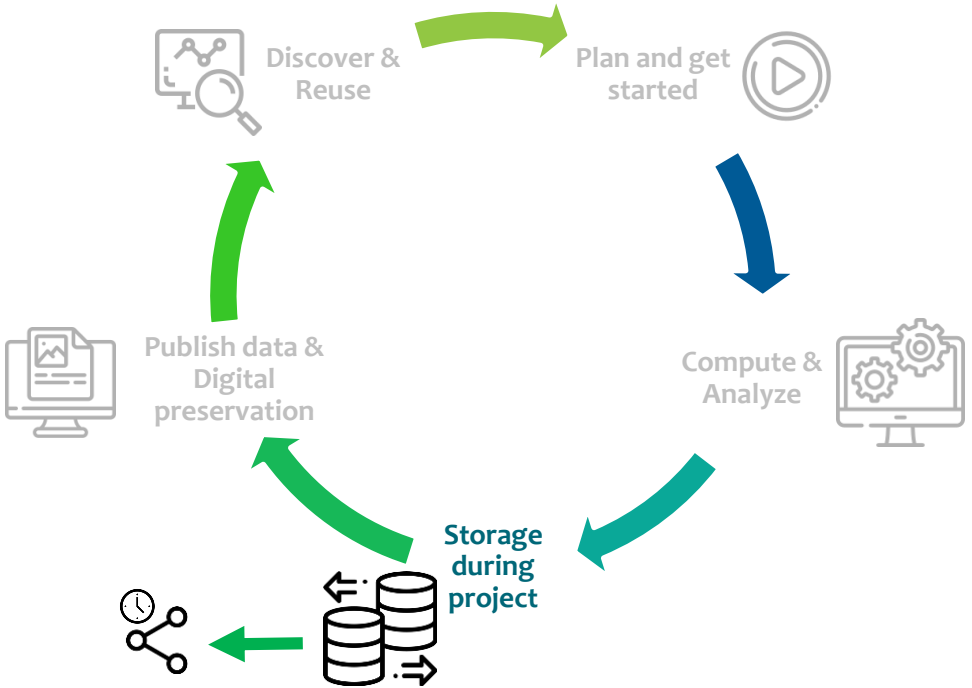


OPEN ASIA
UNIFYING SCIENCE, EMPOWERING INNOVATION
Co-funded by the Erasmus+ Programme of the European Union


Researcher
The researcher produces descriptive information about the data.
Information via APIs:
API
Researcher
Organization
Project
Funding
Infrastructures & services
Metadata will be added through out the project.
Images: flaticon.com
Organisation
Support services: Up-to-date information about services and data in information systems
DMP approval
Impact assessment
Privacy notice
Agreements
National infrastructure
Storage services
Analysis services
Sensitive data services
Discipline specific repository
Data publication DOI
Licenses
DMP
Updated version
Final version
Organisations
People
Datasets
DMP PID
Software
Funders
Publications
Funder
Version 1.0
Version 2.0
Reporting
OPEN ASIA
Co-funded by the Erasmus+ Programme of the European Union

Allas – Data storage service for research projects

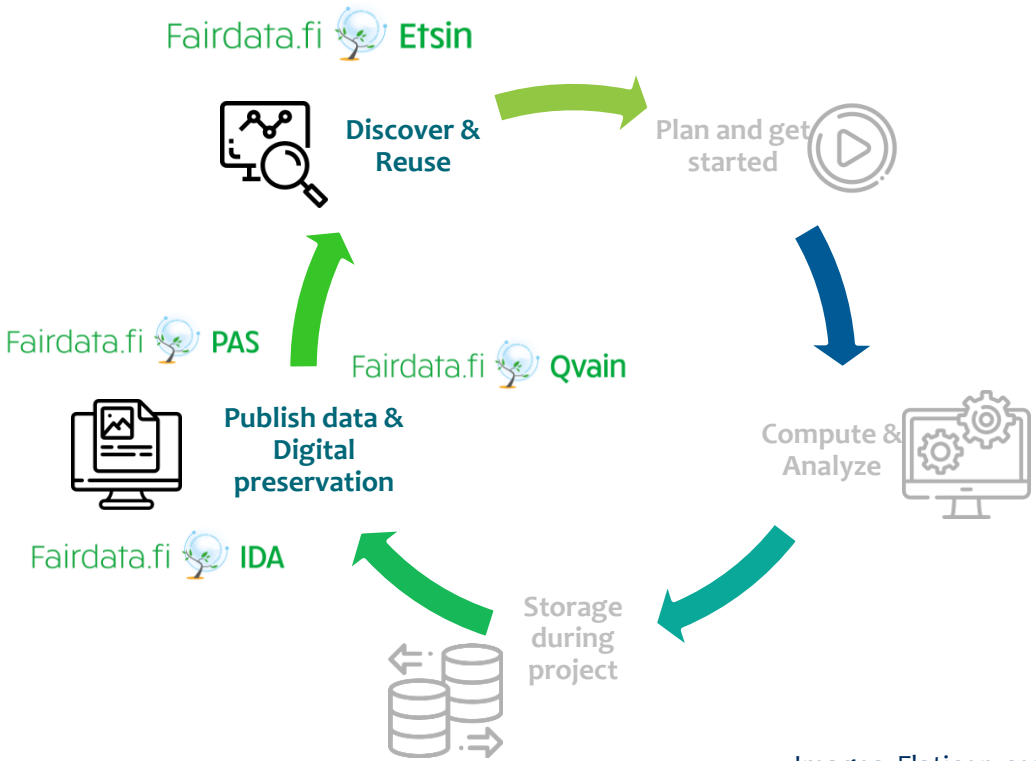
- **Purpose of the service:**
  Storing and sharing research data - especially large amounts of data - during active phase of the research project.
- **Key benefits:**
  - Data processing can be done using standard APIs from anywhere.
  - Data can easily be moved to and from CSC's cloud and computing environments.
  - Free of charge for Finnish academic research
  - Available storage quotas: 10-200 TiB
  - Sensitive data can be stored after encryption
  - Data integrity is maintained with Erasure Coding and checksums
- **More information:** https://docs.csc.fi/data/Allas/



Images: Flaticon.com

Fairdata.fi

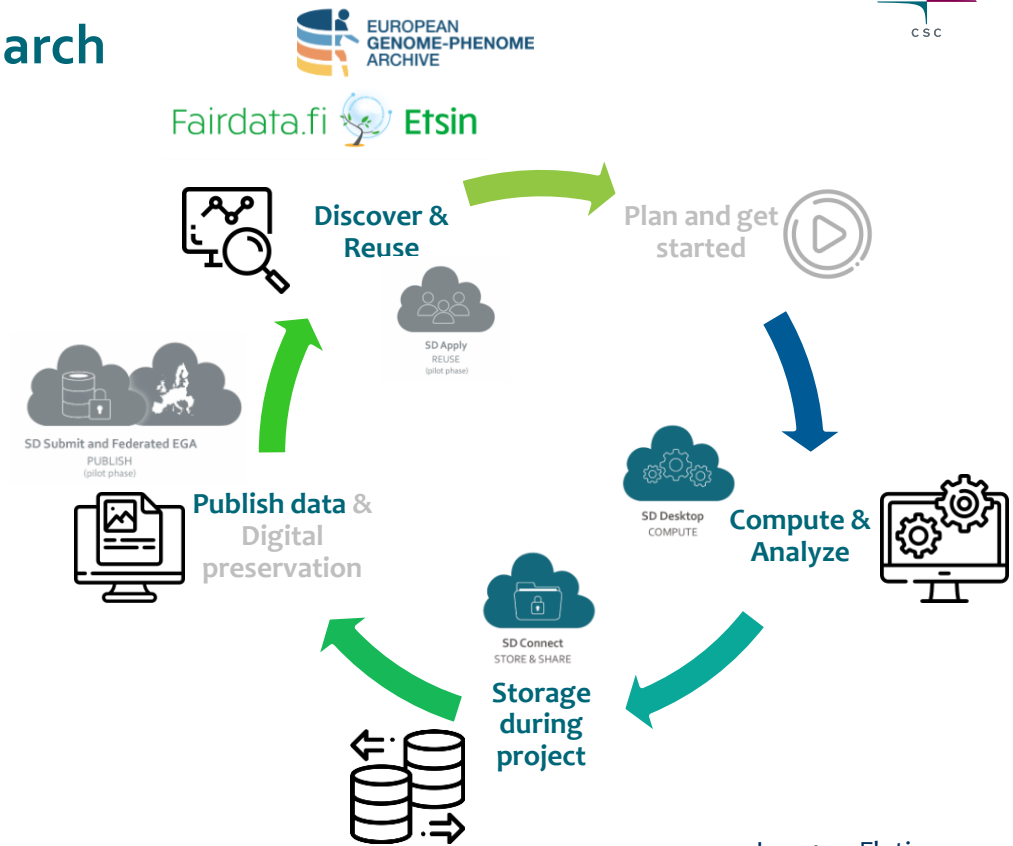
- **Purpose of the service:** Describing, publishing, discovering and preserving research data.
- **Key benefits:**
  - Publishing even large or growing research datasets according to FAIR principles free of charge, with easy to use web tools.
  - Dataset can be transferred to Digital Preservation Service for Research Data.
  - Increased national visibility for published data through research.fi portal.
- **More information:** https://www.fairdata.fi



Images: Flaticon.com
Fairdata.fi

Sensitive Data Services for Research

- **Purpose of the service:** Storing, computing, describing, publishing, and reusing **encrypted sensitive** research data under controlled access.
- **Key benefits:**
  - Available from the public internet, modern web browser (no VPN, no client to install)
  - Data controller has all the tools to manage data access and use (from data collection to reuse).
  - Data are always encrypted (during data storage, data analysis , data export, and re-use)
  - SD Desktop is a certified computing environment for secondary use of health and social data according to Findata regulations.
- **More information:**
  https://research.csc.fi/en/sensitive-data-services-for-research



Images: Flaticon.com
Fairdata.fi

OPEN ASIA
UNIFYING SCIENCE, EMPOWERING INNOVATION
Co-funded by the Erasmus+ Programme of the European Union

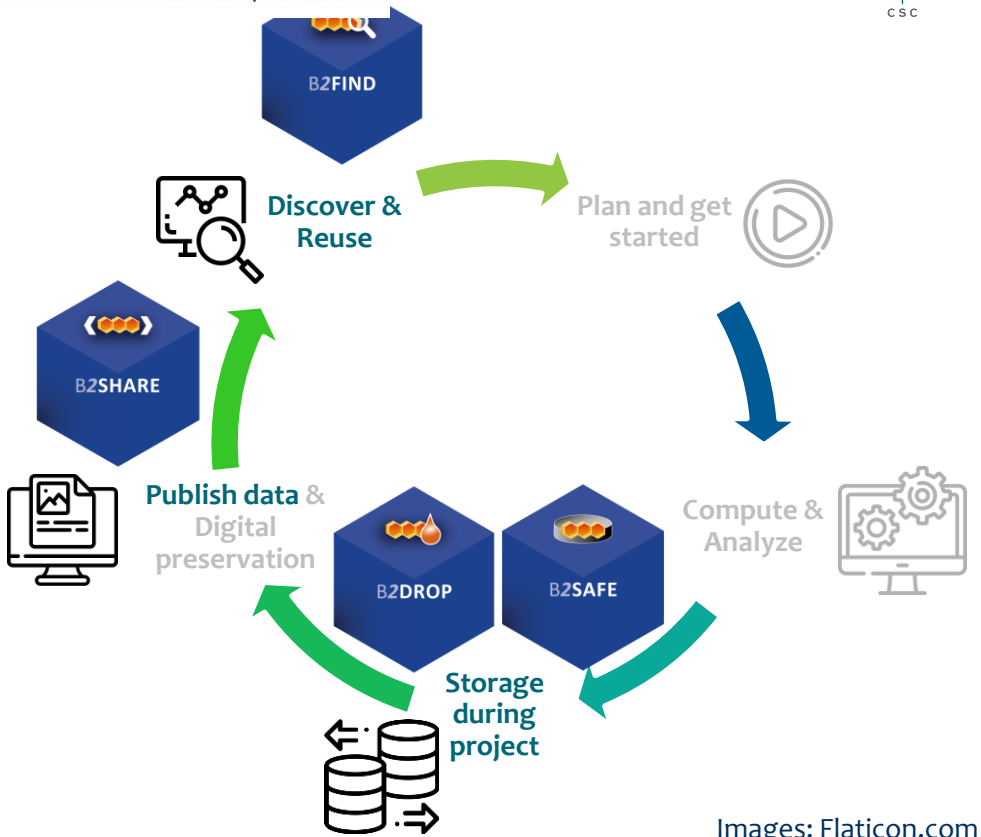
EUDAT Collaborative Data Infrastructure

OPEN ASIA
UNIFYING SCIENCE, EMPOWERING INNOVATION

Co-funded by the Erasmus+ Programme of the European Union

CSC

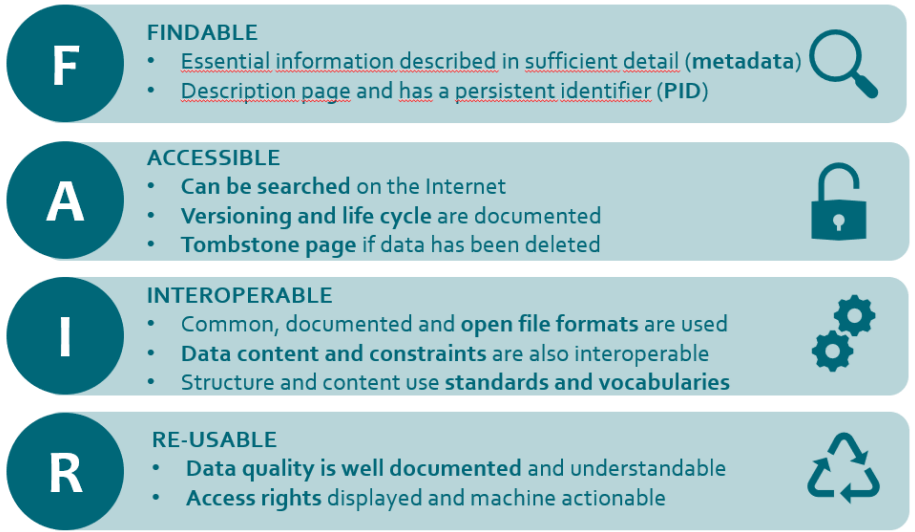
- **Purpose of the service**: Storing, describing, publishing and discovering research data, with an easy access to *international* infrastructures, services and resources.
- **Key benefits:**
  - Can be tailored to meet the specific needs of a community in terms of metadata, storage space, user management and branding.
  - Shared core schema and dataset findability through B2FIND. Support for access across multiple international organizations.
  - Supports policy driven data management through automation and open APIs
- **More information:**
  https://research.csc.fi/-/eudat-services



Images: Flaticon.com
Eudat.eu

How Fairdata services, Sensitive Data Services and EUDAT services can help to comply with FAIR principles

- Services utilize many **common vocabularies** (F, I)
- Stored data can be made part of datasets, described and published (A, R)
- Published dataset has a persistent identifier (PID) and landing page, through which data can be accessed – publicly available data – or requested for re-use – restricted or sensitive data (F, A, I, R)
- Research data documentation is available in an interoperable, machine-readable format (F, A, I, R)



OPEN ASIA
UNIFYING SCIENCE, EMPOWERING INNOVATION

Co-funded by the Erasmus+ Programme of the European Union

Service Comparison Table 1/2

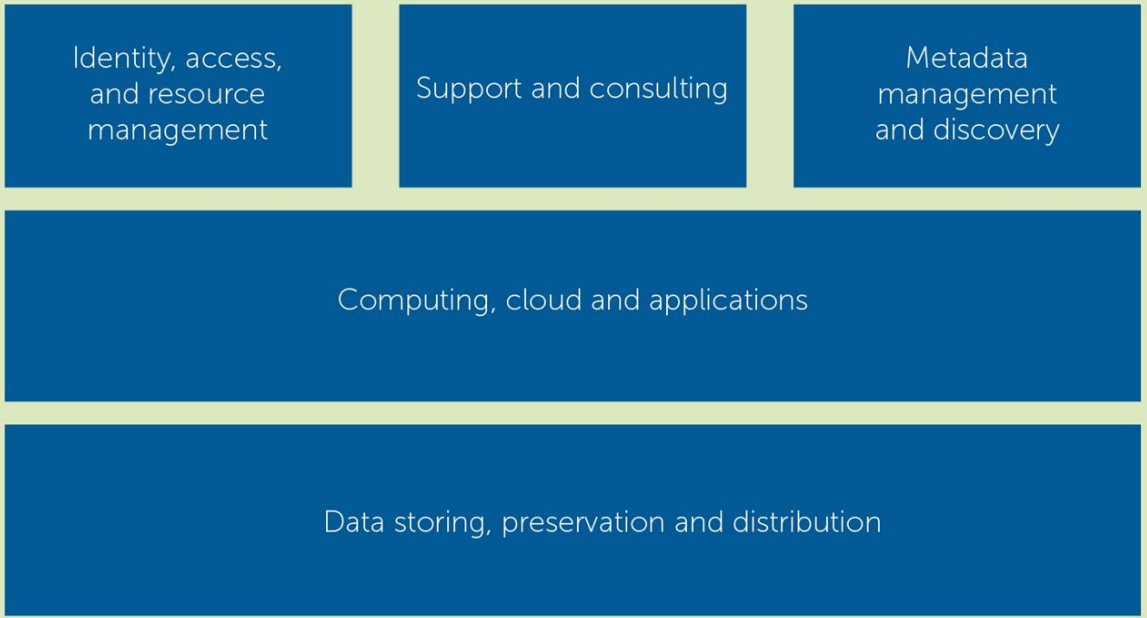
	Fairdata	Sensitive Data Services	EUDAT
Where is the data stored?	In Finland	In Finland	In Finland (replicas in EU, if customer wishes)
Chargeability	• Free of charge for Finnish academic research • Users commit to publishing the stored data in Fairdata Etsin service	Free of charge for Finnish academic research	• Public services: free for everyone • Premium services: based on contract with organization
Tailoring possibilities	No	Limited	Yes, per contract or possible EU-project collaboration
Suitable for sensitive data	• No • In Digital Preservation Service: self assessment to be done by the organization	Yes (automatic encryption)	No
Data integrity	• RAID, Checksums, File replicas • In Digital Preservation Service: part of service promise with e.g. active error correction	Checksums and/or File replicas in some services	RAID, Checksums and/or File replicas in some services per contract

Service Comparison Table 2/2

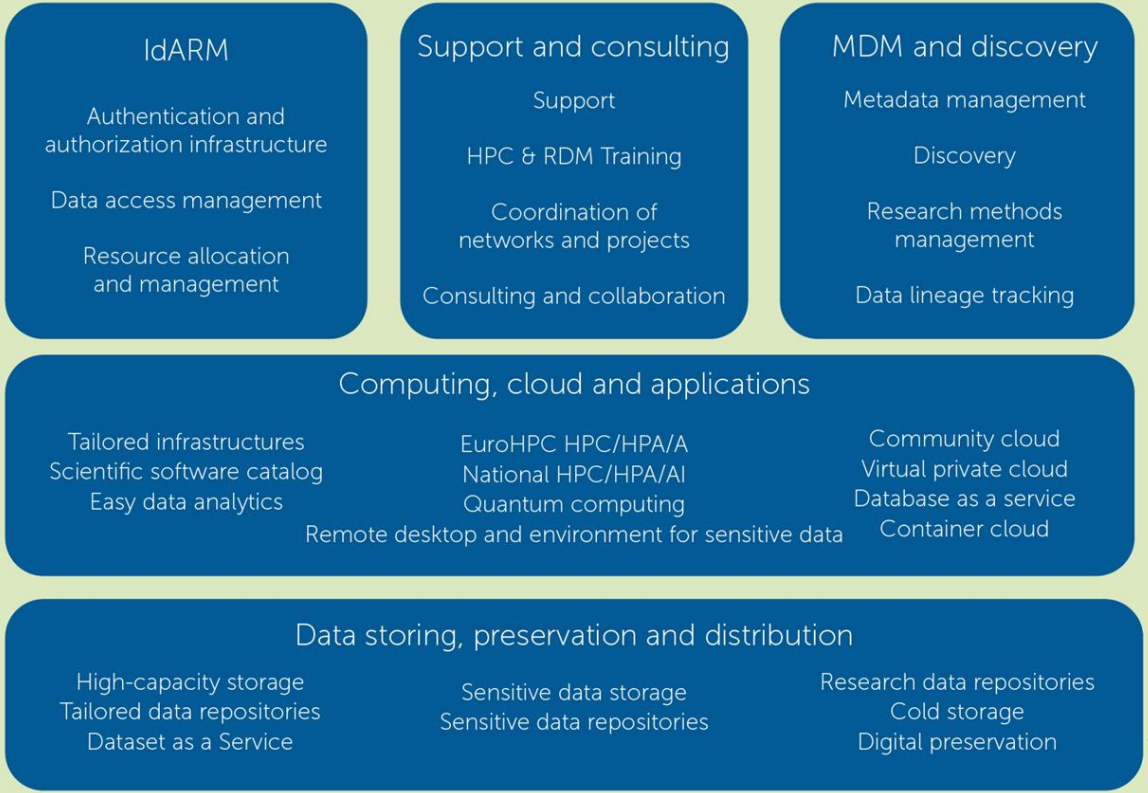
OPEN ASIA
UNIFYING SCIENCE, EMPOWERING INNOVATION
Co-funded by the Erasmus+ Programme of the European Union
CSC

	Fairdata	Sensitive Data Services	EUDAT
Storage period	• Active phase of research project • After the project according to home organization policy • In Digital Preservation Service: decades or even centuries	• Active phase of research project • At the end of the research project data can be published for reuse under controlled access	• Public services: data is stored continuously • Premium services: policy definable by customer
Available storage quotas	• 1 Gib to several TiBs • In Digital Preservation Service: according to Ministry's decisions	• Free academic use default for a new project is 10 TiBs • The quota can be increased with application and agreement	• Public services: 20 GiB per dataset/file/user • Premium services: policy definable by customer
Sharing data	Yes, also dataset metadata can be co-edited	Yes (encrypted via SD Connect or via data streaming using SD Desktop)	Yes
Publishing datasets with persistent identifiers (PIDs)	Yes	Yes	Yes
Curation by service provider	• No • In Digital Preservation Service: keeping data readable and usable, e.g. conversion of obsolete file formats	No	No * bit level (B2SHARE, B2SAFE)

Services for research enterprise architecture
Service areas 2026



Services for research business service architecture



Generally about the services' usage policies, the roles and responsibilities

- CSC does not assert ownership or any intellectual property rights to data in our services.
- It's recommended to agree upon the rights to research data within the research group and research organization early on.
- It is the data owner's responsibility to decide which service is suitable for the data in question and that the encryption - if needed - is done appropriately.
- The data owner decides on the openness and usage policies for their data.

OPEN ASIA
UNIFYING SCIENCE, EMPOWERING INNOVATION

Co-funded by the Erasmus+ Programme of the European Union

OPEN ASIA
UNIFYING SCIENCE, EMPOWERING INNOVATION

Co-funded by the Erasmus+ Programme of the European Union

CSC

Thank you!

I am Hanna Koivula and I work at the CSC as a Senior Research Data Management (RDM) Specialist

You can reach me from:
hanna.koivula@csc.fi

facebook.com/CSCfi

twitter.com/CSCfi

youtube.com/CSCfi

linkedin.com/company/csc---it-center-for-science

github.com/CSCfi