

Data Management Services at SURF



Research Data Management

- **Research Data:** “materials generated or collected during the course of conducting research”, by The National Endowment for the Humanities.
- **Data Management:** Actions that contribute to effective **storage, preservation and reuse of data and documentation** throughout the **research lifecycle**.



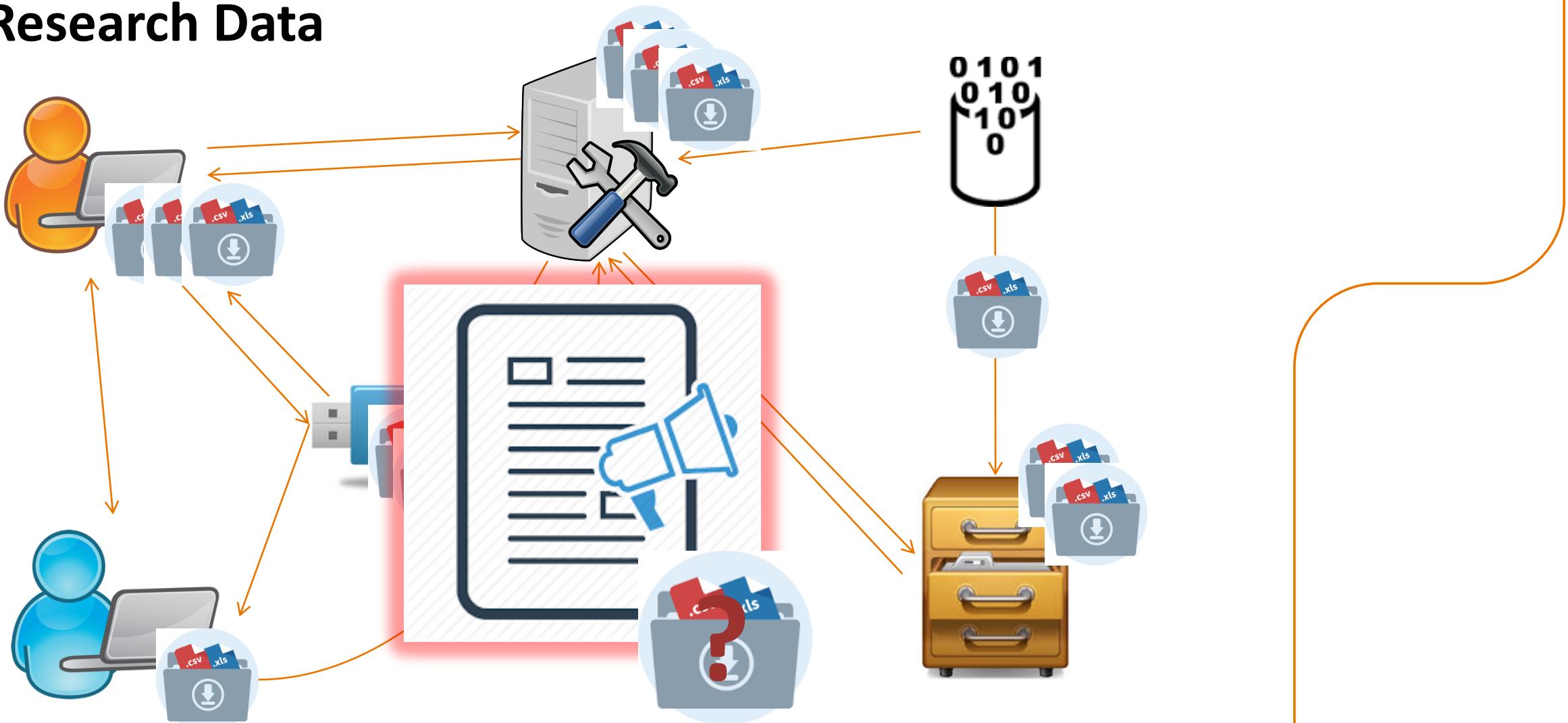
Data Challenges

- Data Explosion
 - More and more data is being created
 - Navigate, transfer and share data
 - Reuse existing data

- Data Loss:
 - Natural disaster, infrastructure failure, storage failure, ...
 - Application software failure, Format obsolescence,...
 - Human error, malicious attack,..



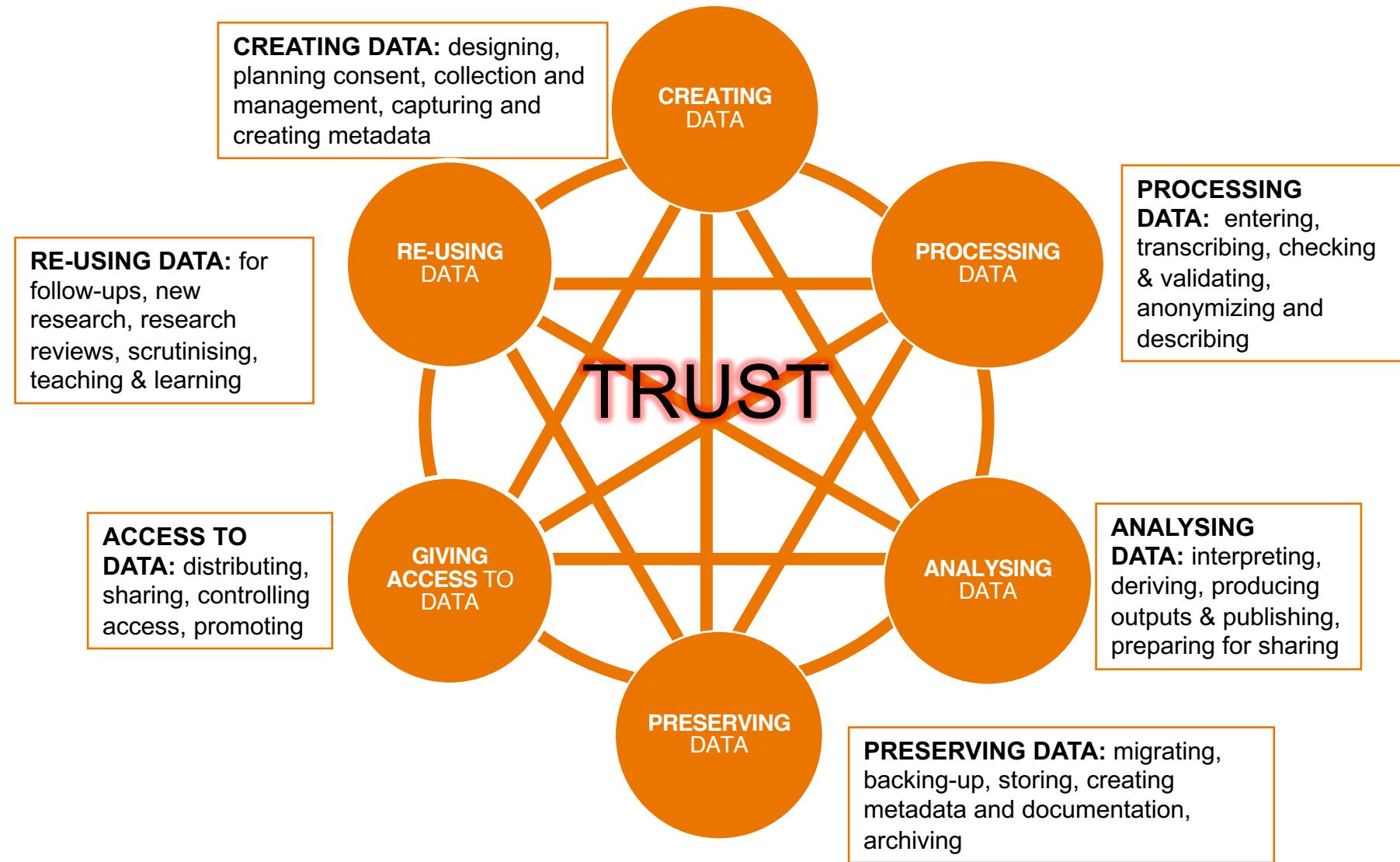
A Typical Problem in handling Research Data



Researcher's needs

- **Store** data during research
- **Share** data during and after research
- **Synchronize** data across different locations
- **Backup** data
- **Archive** data
- **Publish** data
- **Link** publication to processed and raw data
- **Find** data and **make data findable** by others
- Data **transfer**
- Data **provenance**: what happened with the data
- ...

Research Data Lifecycle



Metadata, data about data

What is metadata?

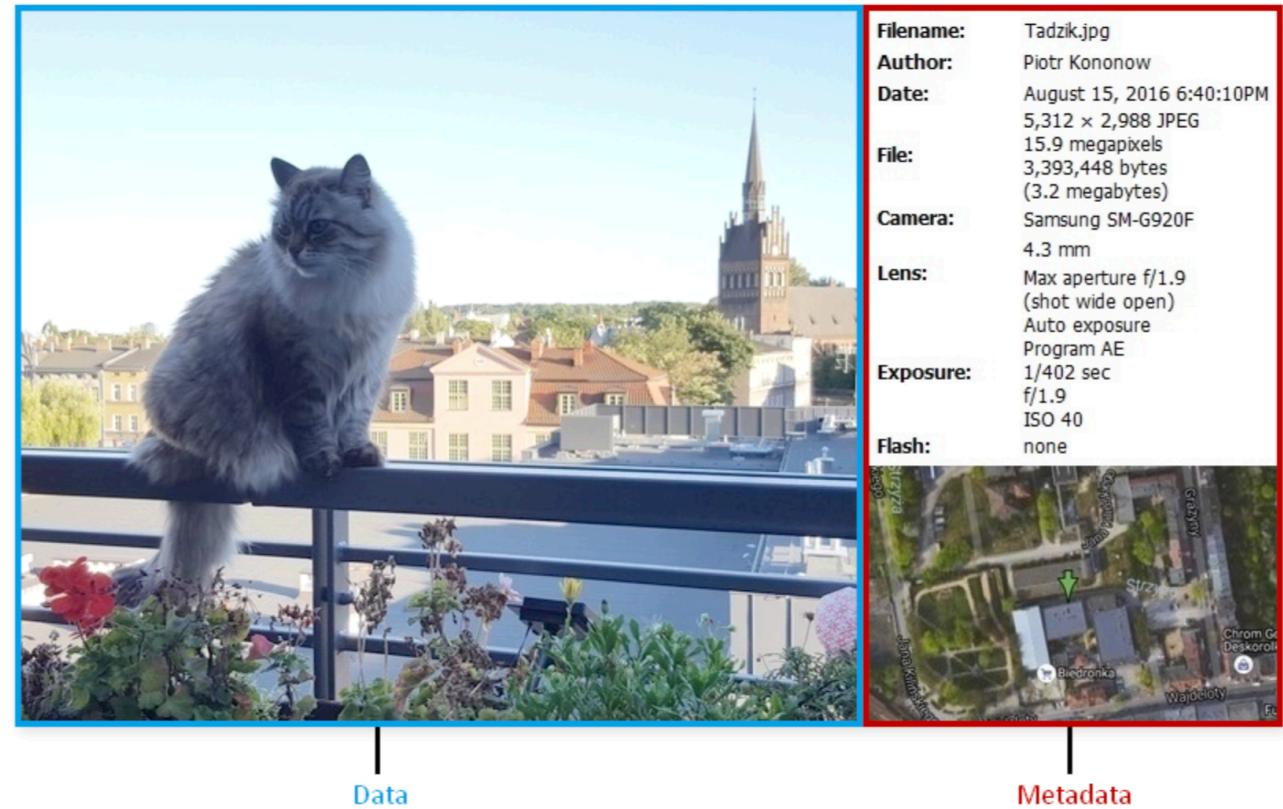
- Metadata is ‘data about data’, which helps to make data findable and understandable

Why using metadata?

- Facilitate data discovery
- Help users determine the applicability of the data
- Enable interpretation and reuse of data
- Allow any limitations to be understood
- Clarify ownership and restrictions on reuse

Types of metadata:

- Descriptive:** information about the content and context of the data
- Structural:** information about the structure of the data
- Administrative:** information about the file type, rights management and preservation processes





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;

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;

Accessible:

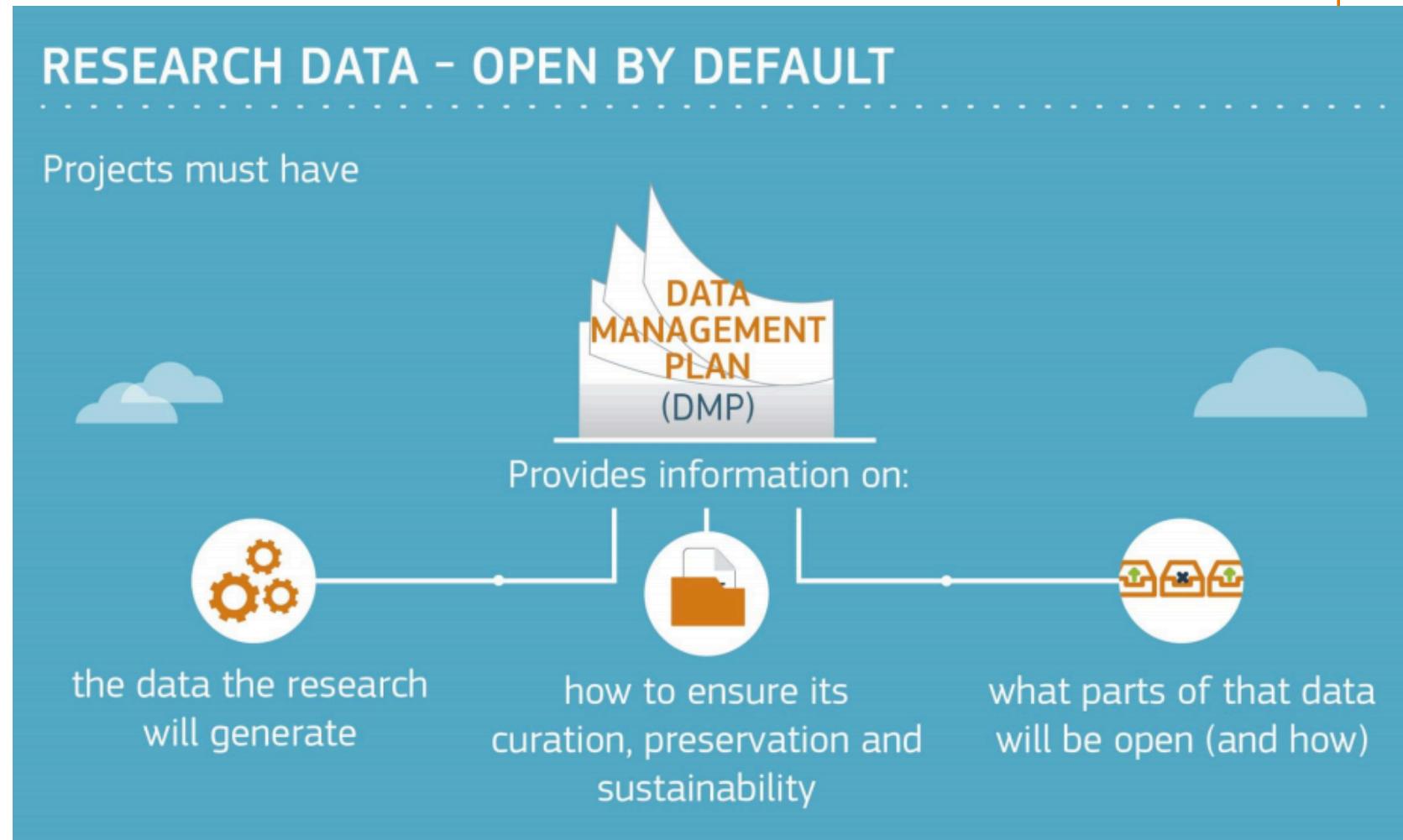
- A1.** (meta)data are retrievable by their identifier using a standardized communications protocol;
 - A1.1** the protocol is open, free, and universally implementable;
 - A1.2.** the protocol allows for an authentication and authorization procedure, where necessary;
- A2.** metadata are accessible, even when the data are no longer available;

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;

Data Management Plan

- **Data Management Plan (DMP):**
A document that outlines how data are to be handled both during and after a research project
 - research funders mandate writing a DMP
 - Type of data
 - Data & metadata standards
 - Data storage, sharing, & preserving
 - Budget & finance
 - ...



SURF's Role in Research

- Vision: Driving ICT innovation in education and research together

Focus of services:

- High Performance Computing (Lisa, Cartesius, HPC Cloud)
- Distributed computing (Grid)
- Big Data
- ...
- → Largest infrastructure for (storing & processing) large data



SURF Data Services



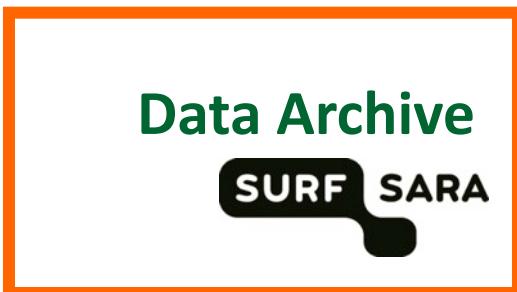
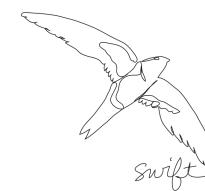
Handle.Net®



SURF FILESENDER

SURF DRIVE

SURF object storage



HPC and Data Management

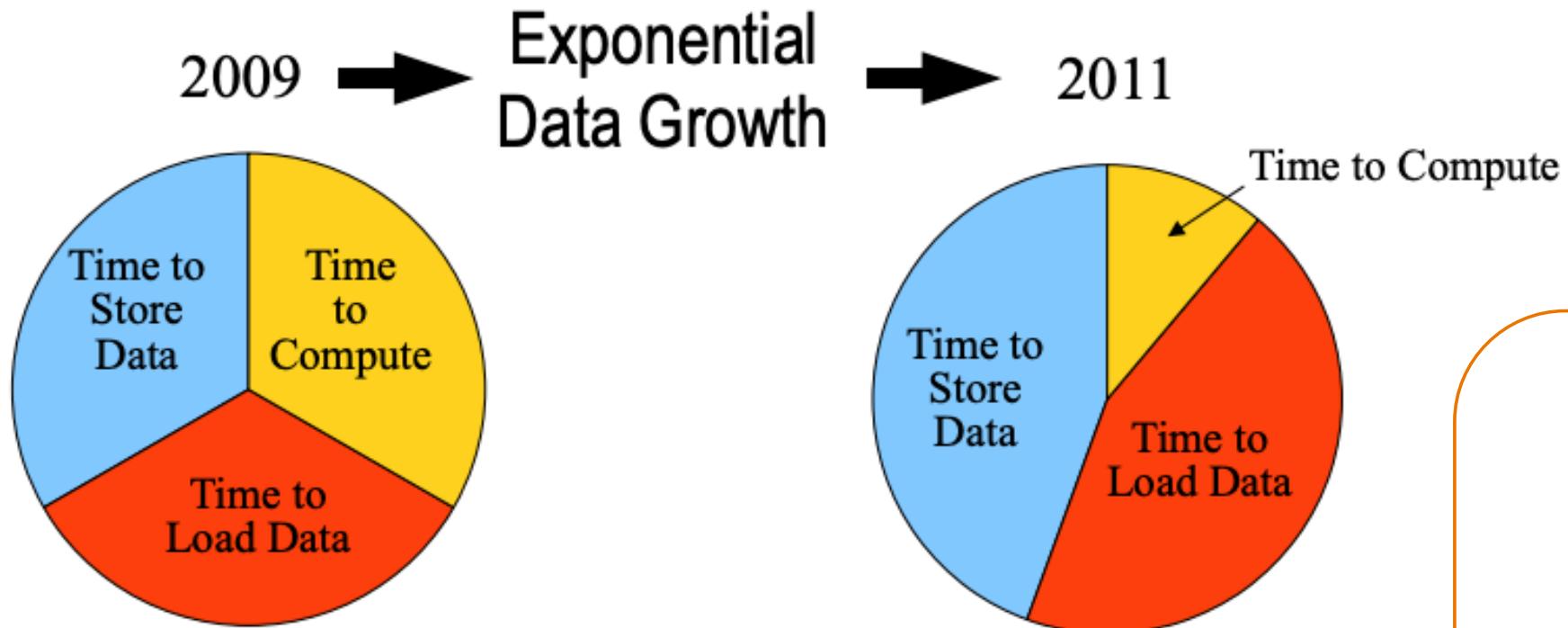
Importance of HPC

- Computationally intensive applications
- Need to analyze more and more data
- Reduce costs and increase efficiency

Challenges/Barriers to HPC

- Data access/sharing
- Complexity of use
- I/O bottlenecks
 - Leads to poor overall application performance
 - Prevent applications from scaling

HPC and Data Access

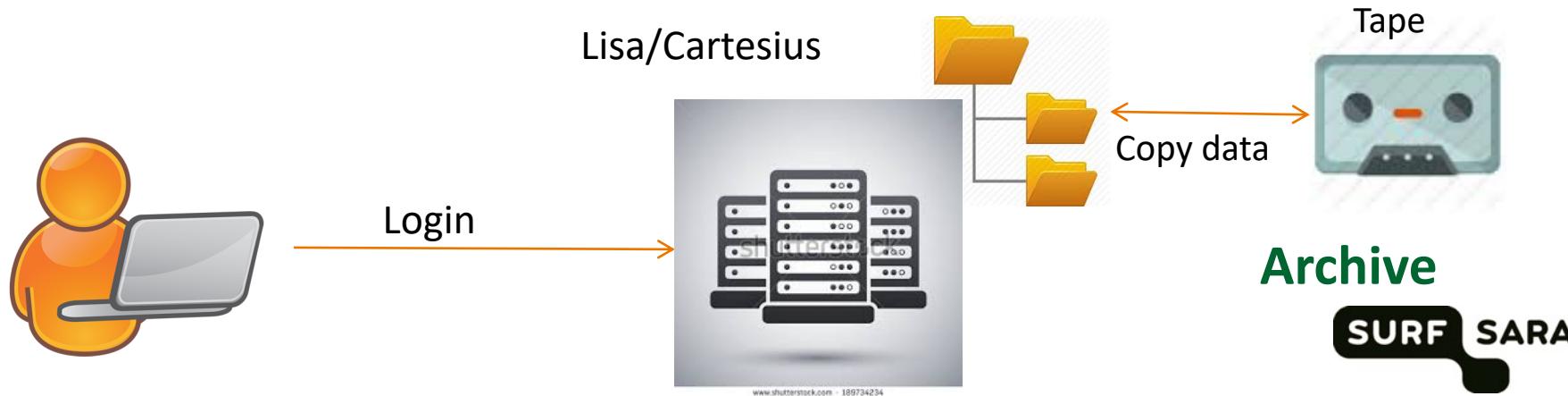


You can only compute as fast as you can move the data

Archiving Data in HPC: Data Archive Service

- Long-term storage of big data
- Storage medium: Tape → high latency

- Powerful transfer protocols:
 - gridFTP, rsync, scp,...



- Easy access from HPC services Lisa and Cartesius via NFS mounts
→ use archive as yet another directory

Data growth in the Data Archive: 10s of Petabyte of data

Figure 4: Usage overview: average file size per domain (in GB)

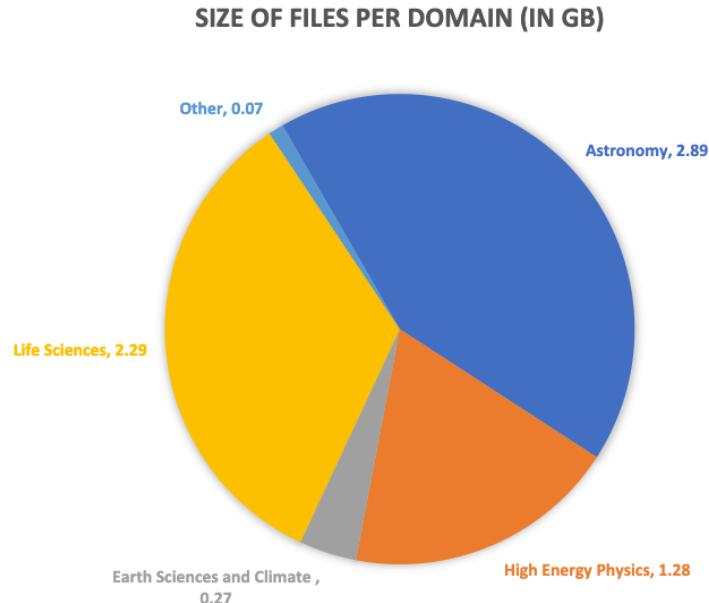
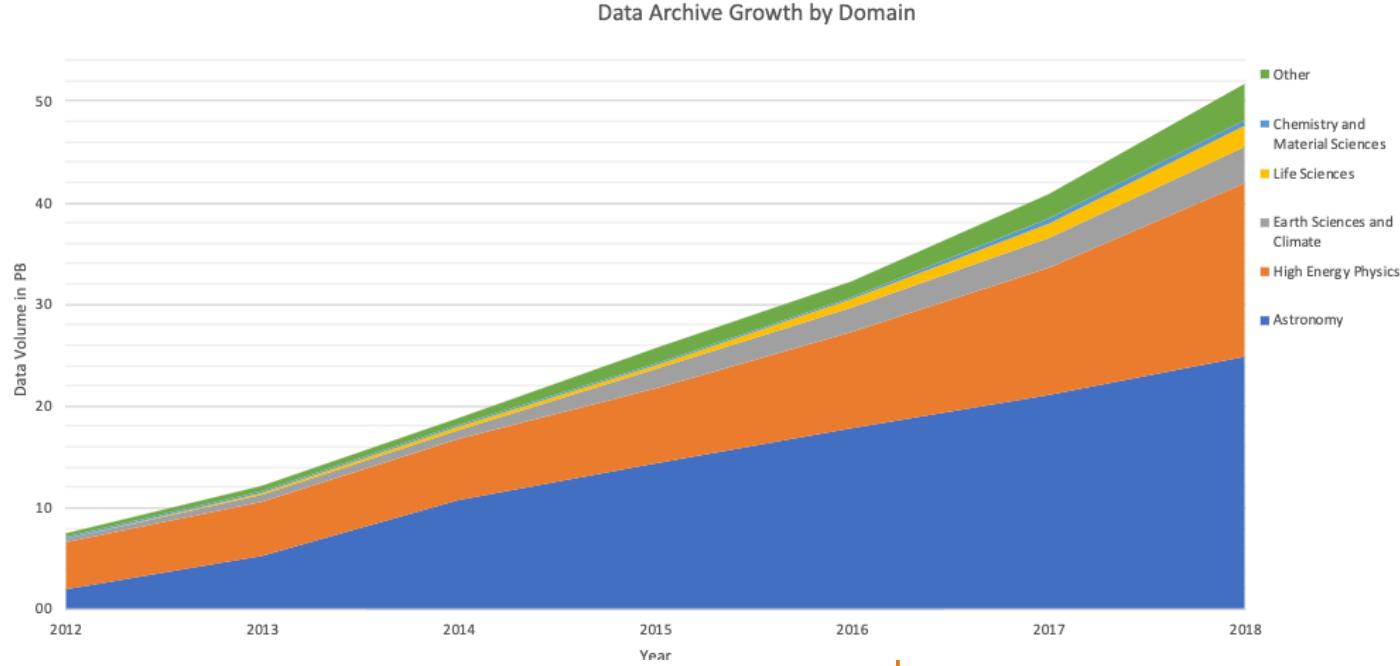
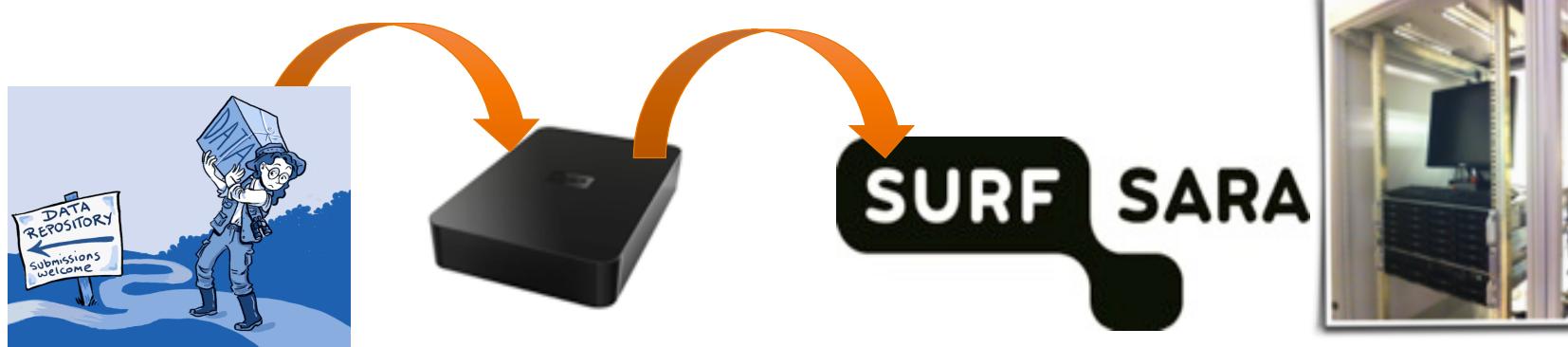


Figure 5: Data on the Archive over time per scientific domain (in PB)

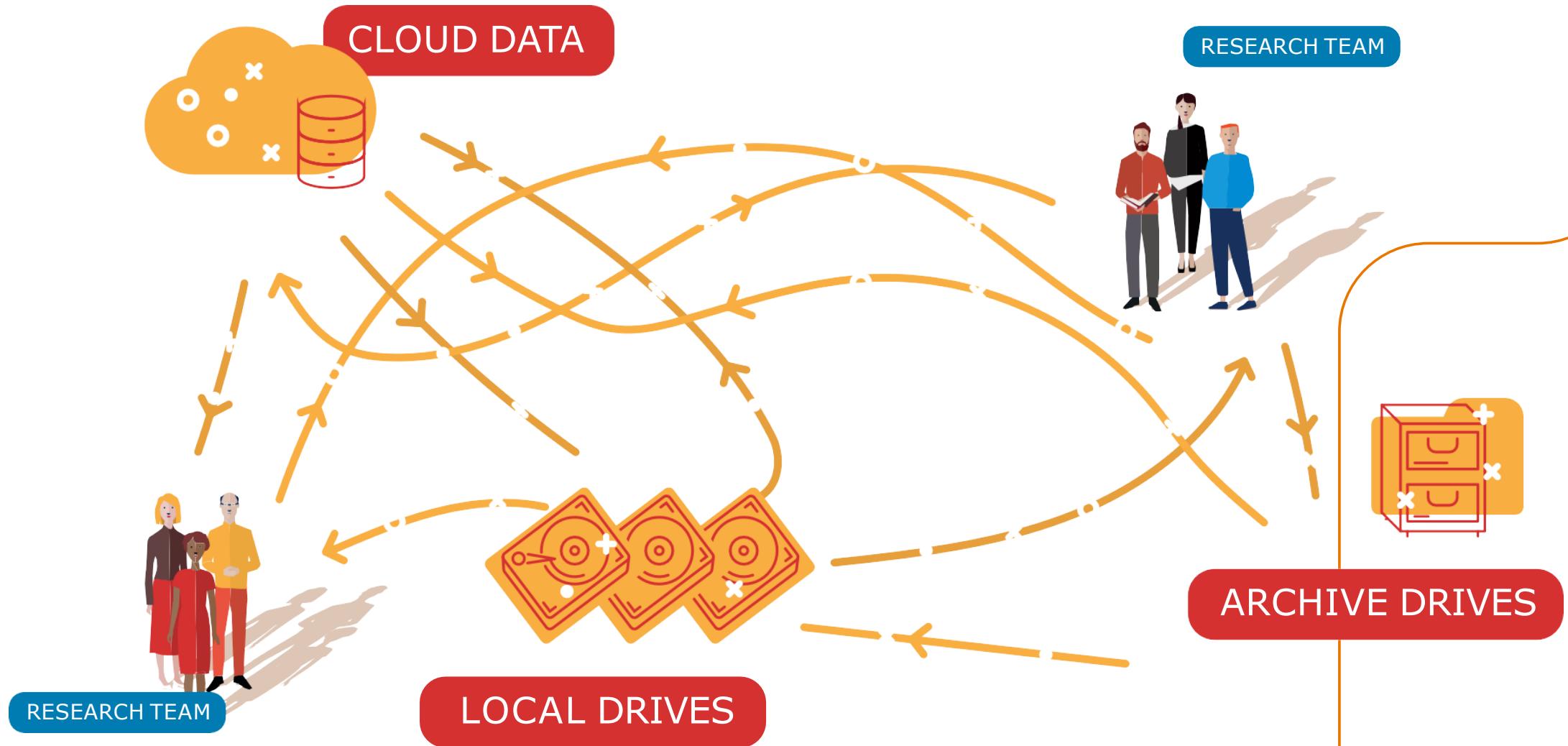


Data Ingest Service

- Data often resides on external storage media, USB sticks, external hard drives
- Slow or no internet connection
- Easy way to upload large data from disk to SURFsara facilities
- Upload data from 45 disks in parallel



Difficulty in Sharing Data



Data Sharing and personal storage: SURFdrive

- Trusted community cloud for personal storage
- Collaboration between SURFsara, SURFnet and Dutch universities
- Specifications and service determined by end-users (universities)
- Sharing smaller data files, documents,...
- 250 GB storage capacity per user
- Privacy and security: Data in the Netherlands
- Based on ownCloud,
- Synchronizes with local storage,...
- Access through: surfdrive.nl

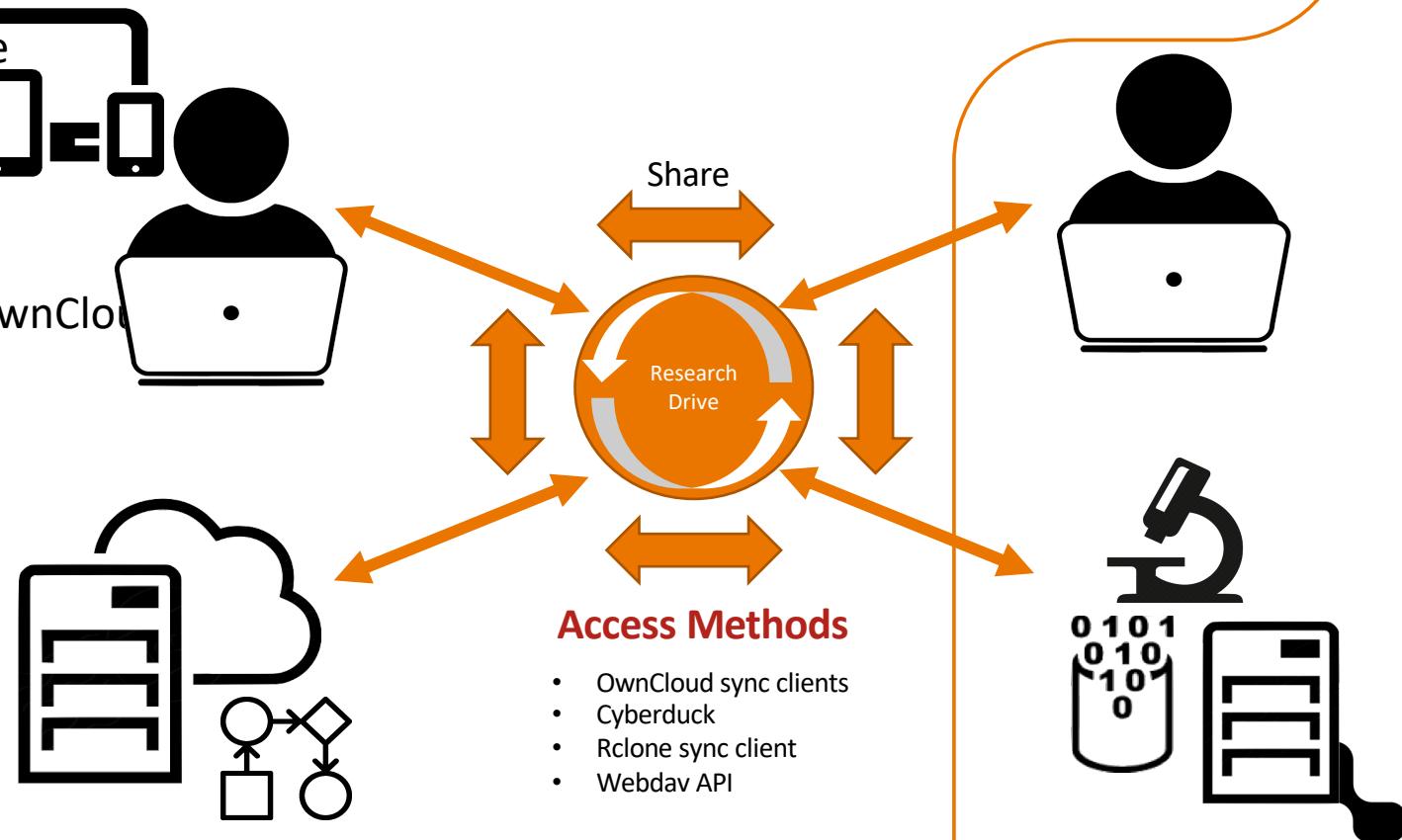


Data Sharing and Team storage: ResearchDrive



COLLABORATIVE WORKING AND SHARING OF DATA FOR RESEARCH COLLABORATIONS

- ✓ Trusted cloud storage for research teams
- ✓ Facilitates collaborative research between teams
- ✓ High quota's (>250 GB) per user/group available
- ✓ Sharing across similar cloud storage services
- ✓ Based on ownCloud
- ✓ Synchronization with local storages, Based on ownCloud
- ✓ Data stewardship



Publishing Research Data: SURF Data Repository

- Data repository service to deposit and publish data
- Long-term preservation of research data
- Connected to tape and Data Archive in the backend
- Provides quality to data sets via metadata descriptions
- Makes data citable and findable via Persistent Identifiers
- Status: Pre-production phase



PID Service

PIDs (Persistent Identifiers) ensures the findability of your data

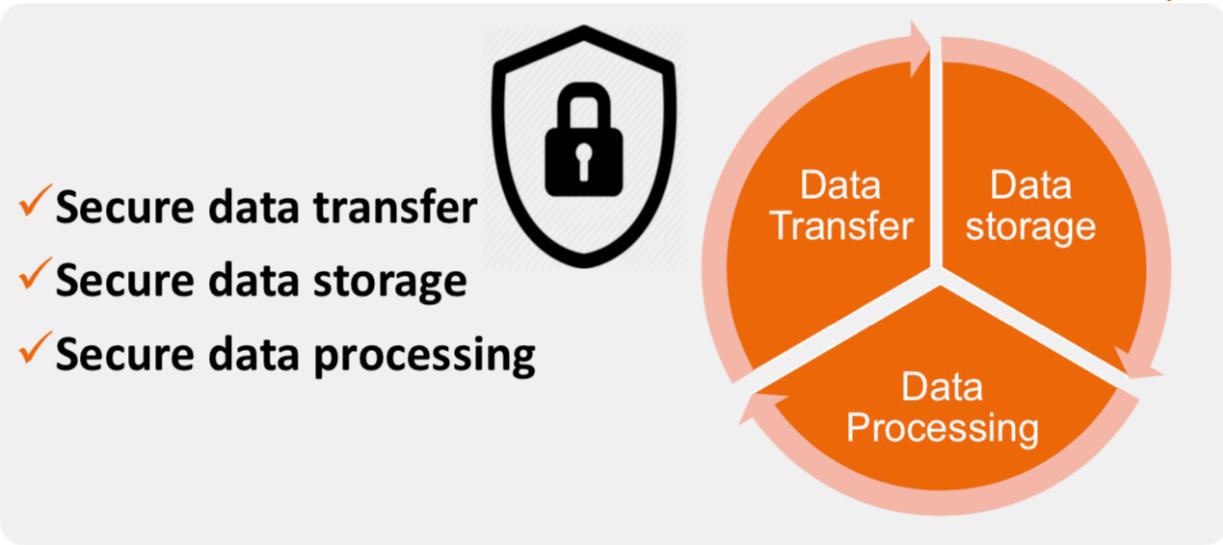
- Pointers to resources like files, folders, webpages, real world objects
- Globally unique and resolvable via http
- Comparable to ISBN numbers assigned to books



- Example resolvers: <https://dx.doi.org/> and <http://hdl.handle.net/>
- A PID consists of a prefix and a postfix ([10.2307/748467](https://doi.org/10.2307/748467))

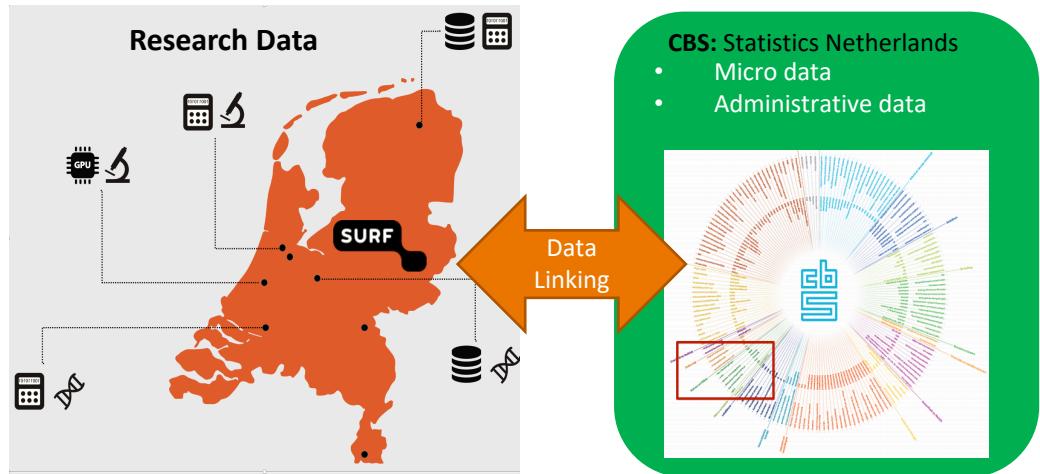
Sensitive Data and HPC

- Sensitive data is data that must be protected against unwanted disclosure
 - Clinical Data, DNA sequences,...
 - Patient data, personal Data
- Provide a secure environment on HPC for sensitive data
- Comply with all legal and ethical requirements
- GDPR, WGBO (Act on the Medical Treatment Agreement), CBS law

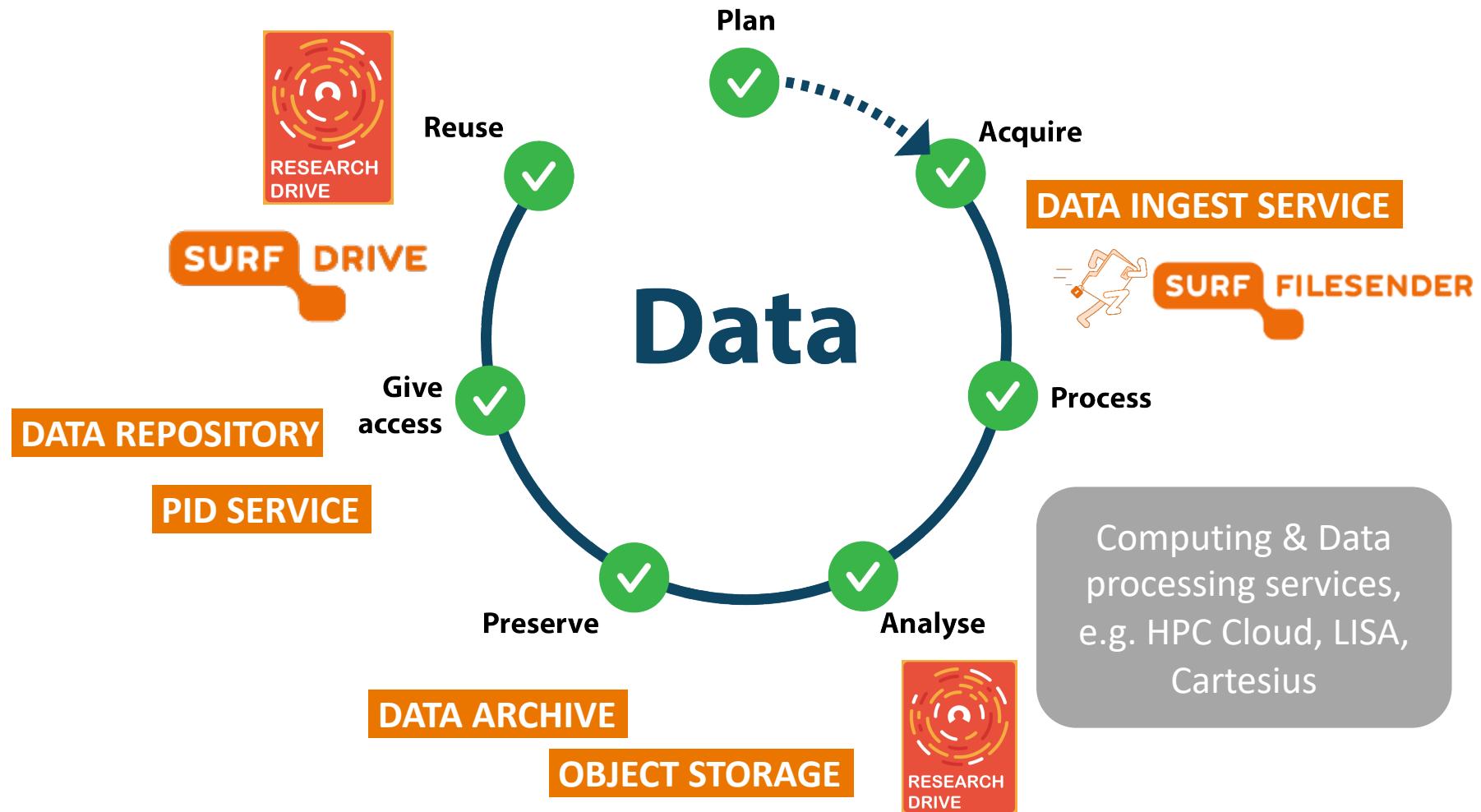


The ODISSEI Data Facility

- A High-Performance Computer environment with secure access to sensitive data
- **Secure sandbox for processing data**



SURF Data Services



Thank you!

Thanks to SURFsara's ...

- ... Data Preservation team
- ... Data Management team
- ... Online Data Services team

