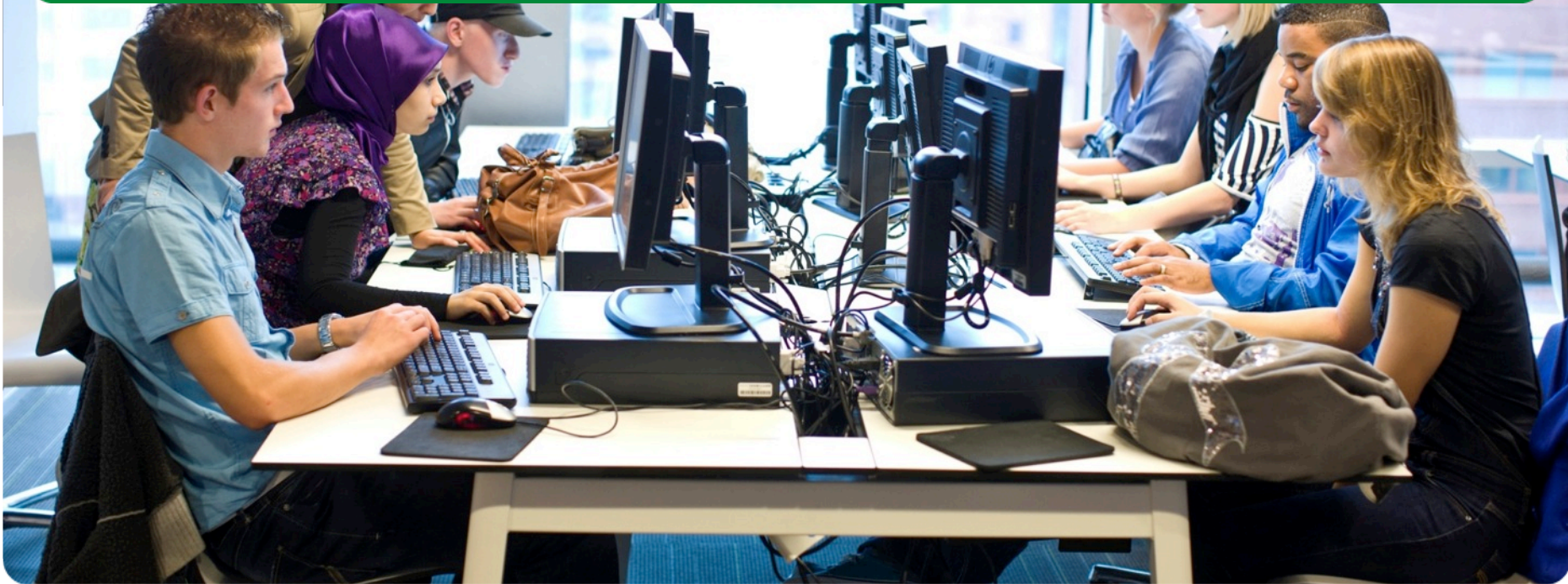


# Data Management Services

NARGES ZARRABI



UvA HPC Course – 29<sup>th</sup> January 2018



# Course Outline

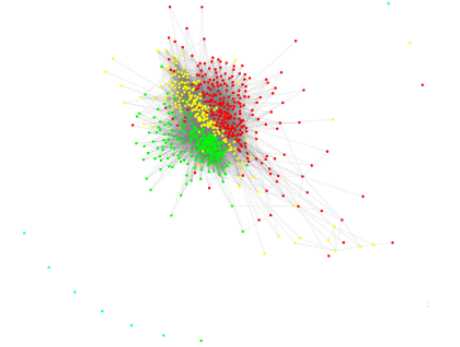
- **Data Management Services at SURFsara (50 min)**  
Demo FileSender
- Break (10 min)
- **Data Archive Infrastructure and Access (50 min)**  
Access Archive via GUI (Demo)  
Access Archive via command line (hands-on)
- Break (10 min)
- **Data Management with iRODS (30 min)**
- **iRODS icommands (hands-on) (30 min)**
- Break ( 10 min)
- **iRODS icommands continued (hands-on) (20 min)**
- **iRODS GUI Demo (30 min)**

# My Background

- Masters in Computational Science at UvA (2008-2009)
- PhD in Computational Science at UvA (2010-2013)



## Complex Networks and Agent-Based Models of HIV Epidemic

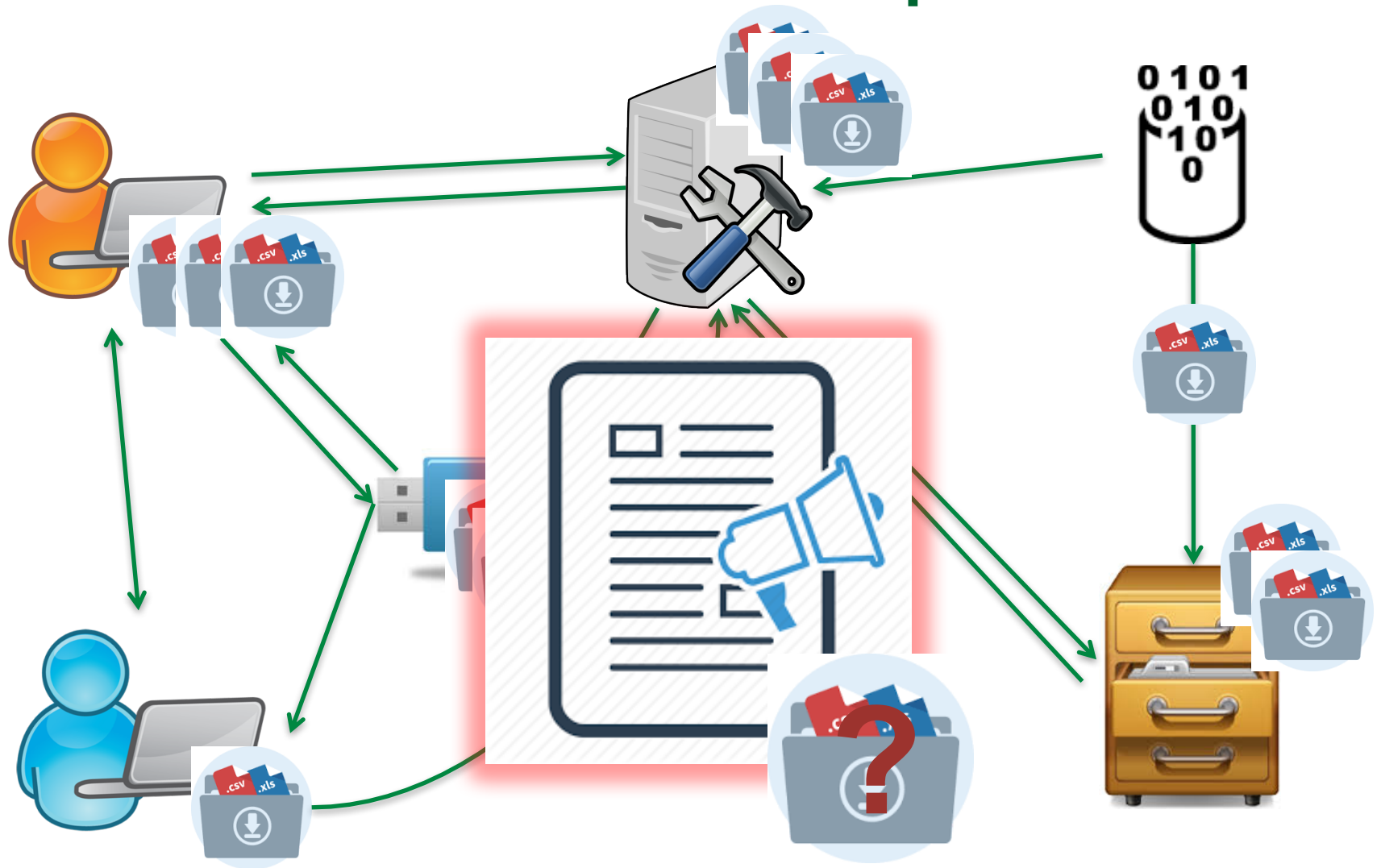


Narges Zarrabi



UNIVERSITY OF AMSTERDAM

# Data – where is the problem?



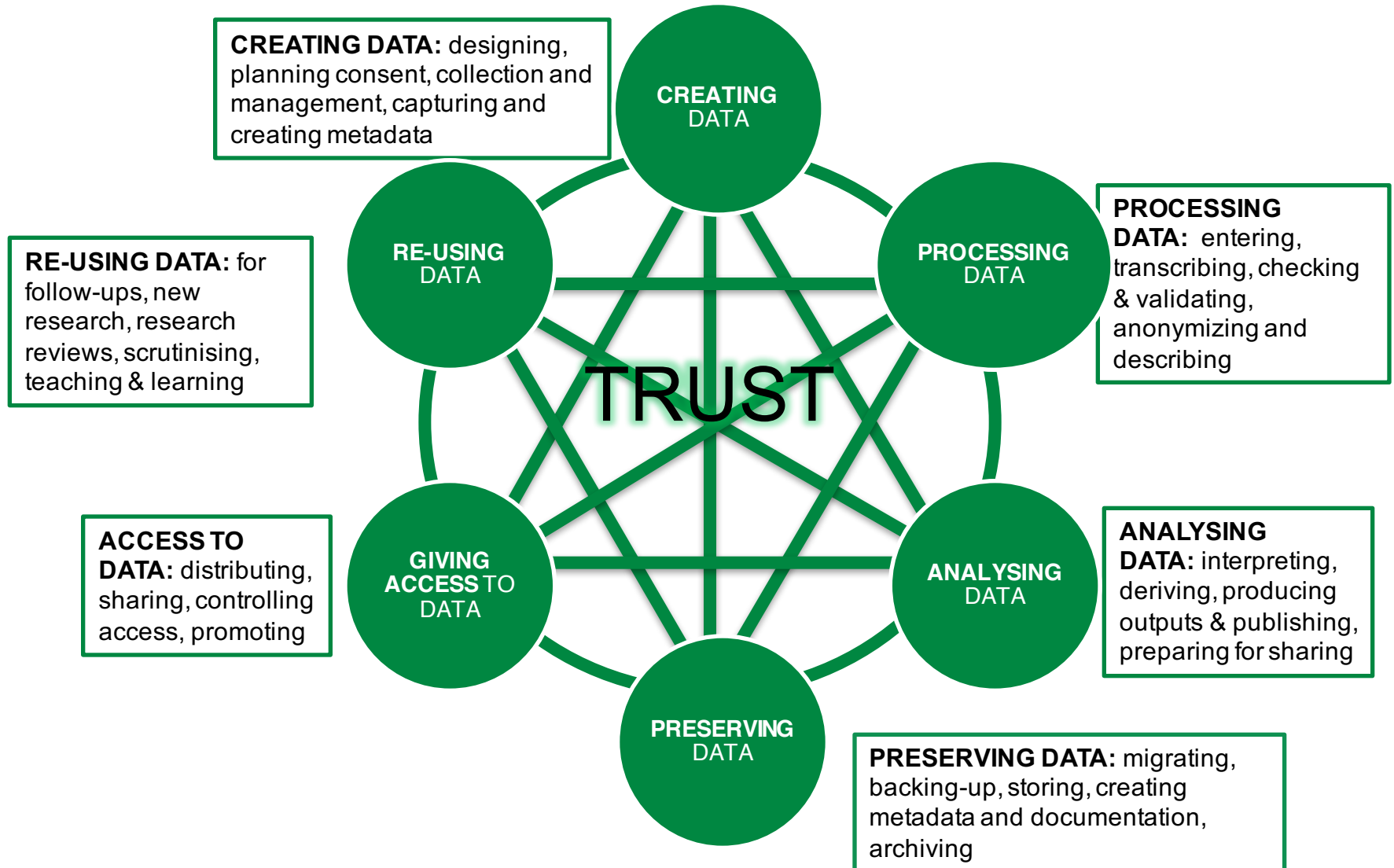
# The researchers' needs

- **Store** data during research
- **Share** data during and after research
- **Synchronise** 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 **transfers**
- Data **provenance**: what happened with the data
- ...

# Data Management

- Actions that contribute to effective **storage**, **preservation** and **reuse of data** and **documentation** throughout the **research lifecycle**.
- **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
  - ☐ Sharing
  - ☐ Transition from collection to reuse

# The data Life cycle





# Data services in the Netherlands – there are a lot of solutions



4TU.Centre for Research Data



Handle.Net®

SWIFT storage service



Archive



Data ingest service



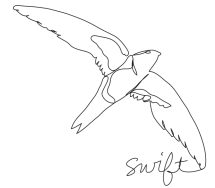
# Data services in the Netherlands – there are a lot of solutions



Handle.Net<sup>®</sup>

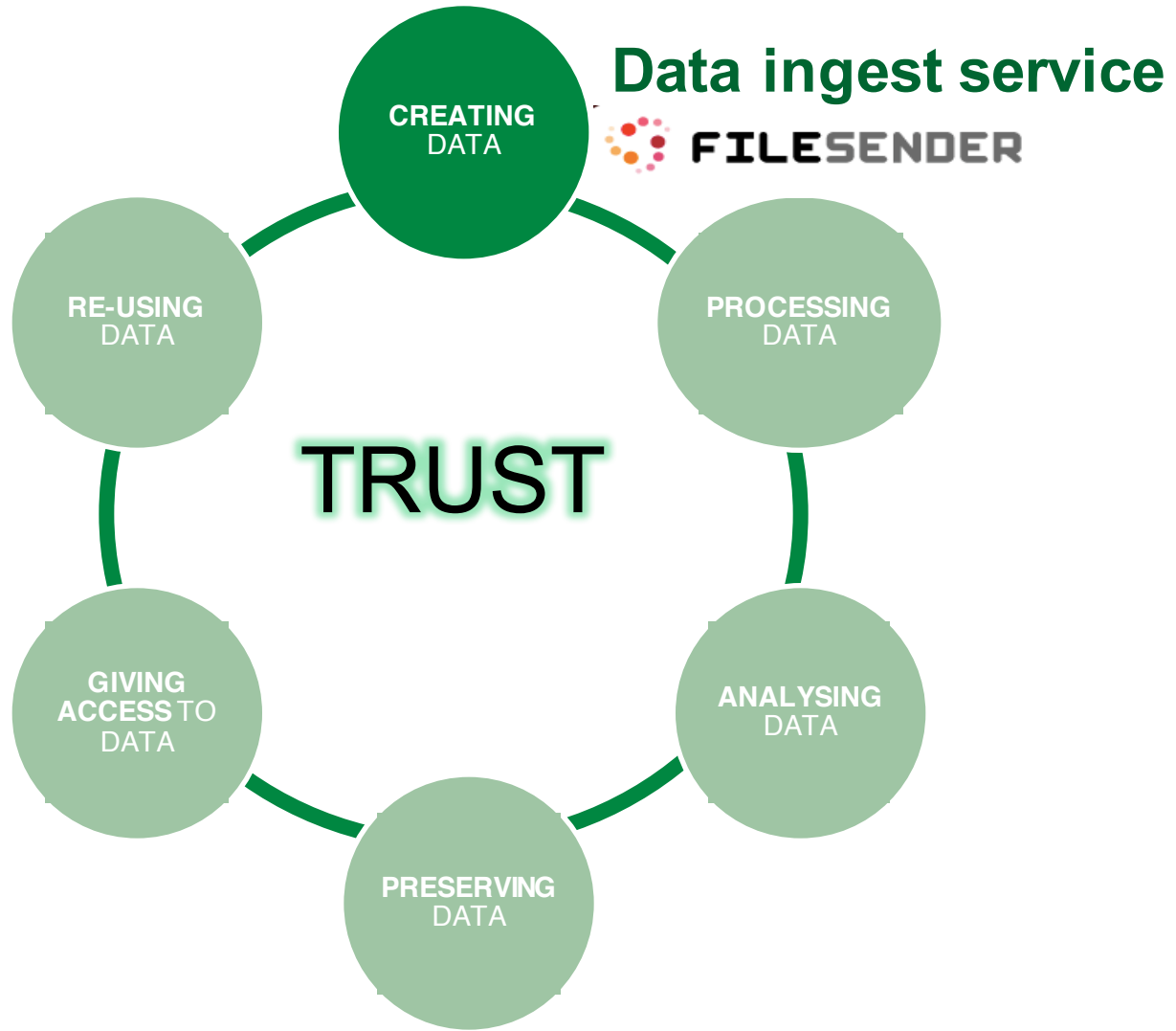


SWIFT storage service



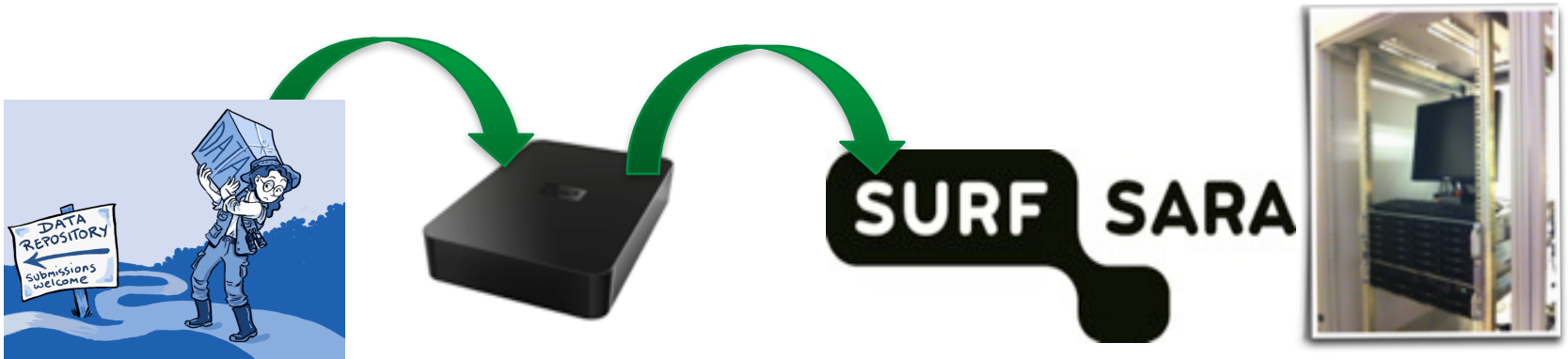
Data ingest service

# The data life cycle



# 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



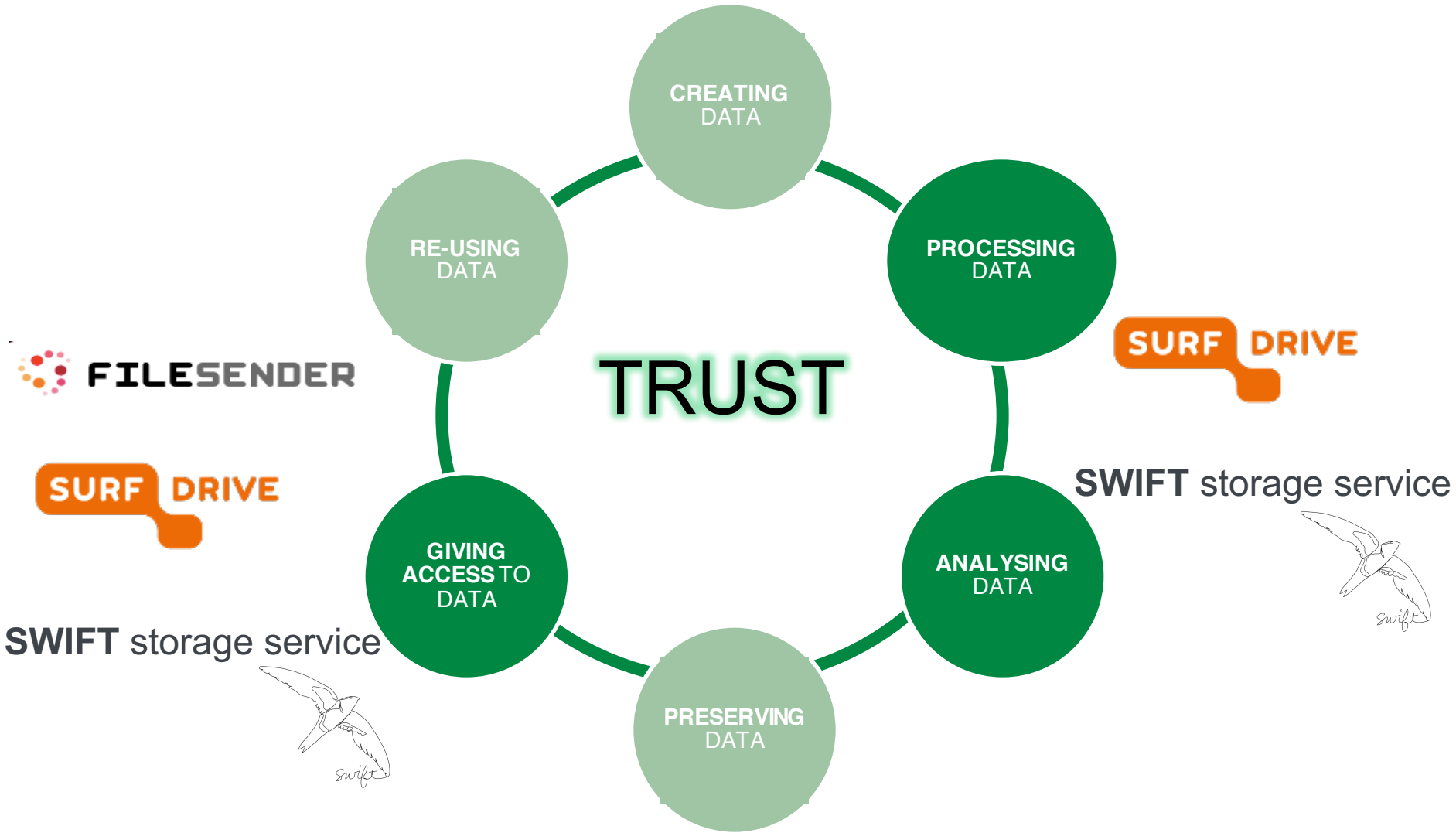
# FILESENDER

**DEMO**

- Trusted community service
- Transferring BIG files from person to person
- File Transport service not File Storage (!)
- Simple interface
- Option to apply end-to-end encryption (250MB browser limit)
- Vouchers for guest usage
- <https://filesender.surfnet.nl/>

The screenshot displays the SURFfilesender web interface. At the top, the title 'SURFfilesender' is visible alongside navigation links for 'Send File', 'Guest Voucher', and 'My Files'. A 'Welcome Paul van Dijk' message is shown on the left, and a '1.6 HTML5' status indicator with a green checkmark is on the right. The main section is titled 'Send a file' and contains a form with the following fields: 'To:' (empty), 'From:' (paul.vandijk@surfnet.nl), 'Subject: (optional)' (empty), and 'Message: (optional)' (empty text area). To the right of the form, a numbered list indicates the steps: 1. Enter delivery email address(es), 2. Set expiry date, 3. Browse for a file, and 4. Click Send. Below the message field, the 'Expiry date:' is set to '27/04/2015'. The 'Select your file:' section shows a 'Choose File' button and 'no file selected'. At the bottom, there is a checkbox for 'I accept the terms and conditions of this service.' with a '[Show/Hide]' link.

# The data Life cycle

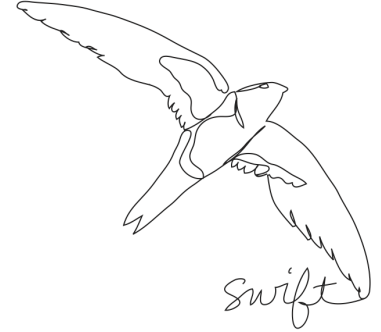


# SURFdrive

- Trusted community cloud for personal storage
- Sharing smaller data files
- Collaboration between SURFsara, SURFnet and Dutch universities
- Specifications and service determined by end-users (universities)
- 250 GB storage capacity per user
- Based on ownCloud, synchronises with local storage
- Access through: [surfdrive.nl](http://surfdrive.nl)



# SWIFT storage service



- Online cloud storage service
- SWIFT is an object storage system
- Ideal for storing various kinds of data that can grow without bound

## Access Methods

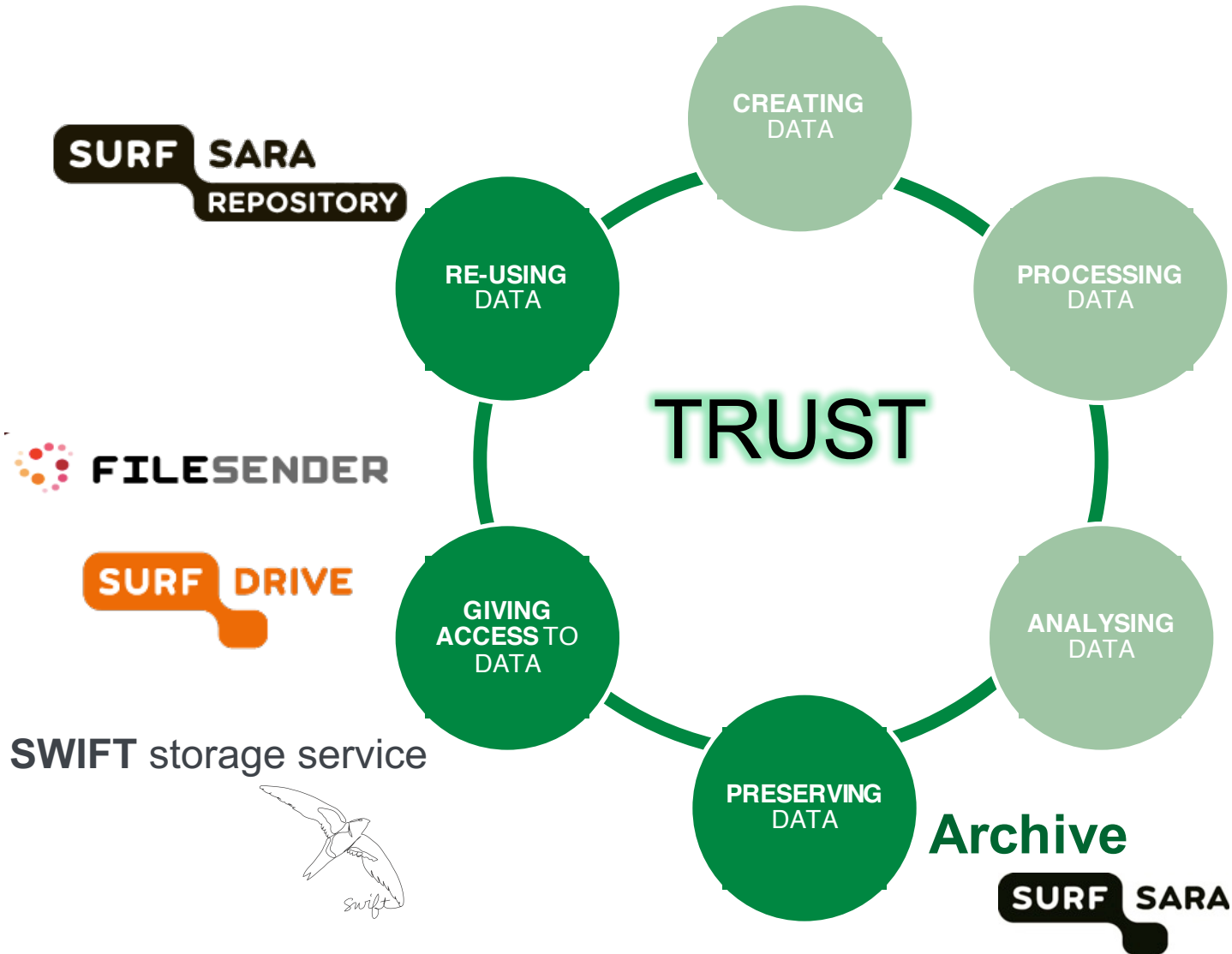
- SWIFT command line client
- Next Cloud
- CURL
- S3 clients
- Cyberduck
- Python library
- API
- ....

## Current Status

- Pre-production phase
- Open for pilot projects.  
Contact: [helpdesk@surfsara.nl](mailto:helpdesk@surfsara.nl)
- Documentation:  
<https://doc.swift.surfsara.nl/en/latest/>

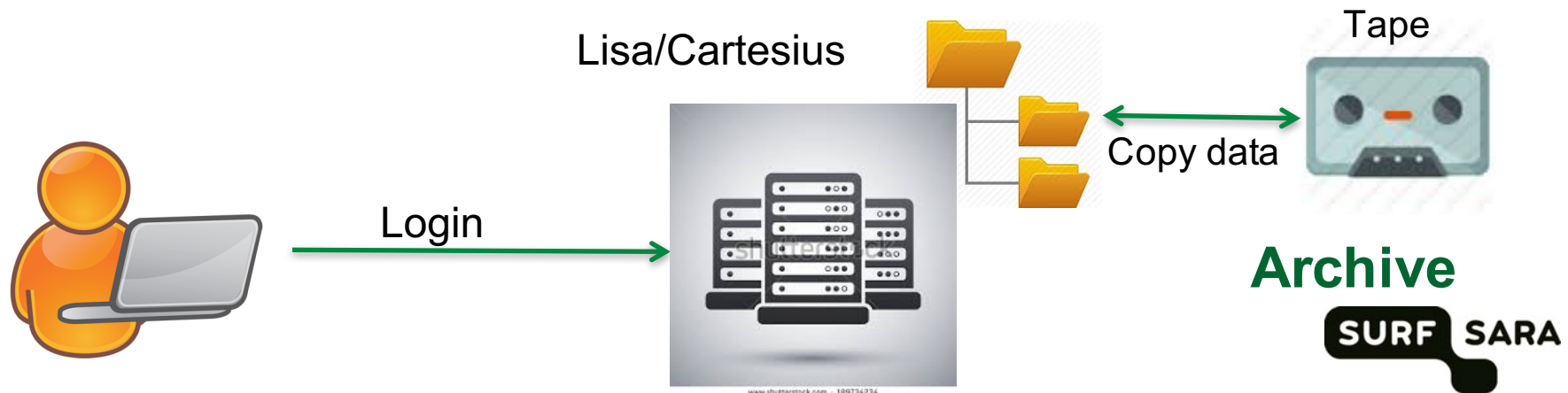


# The data Life cycle



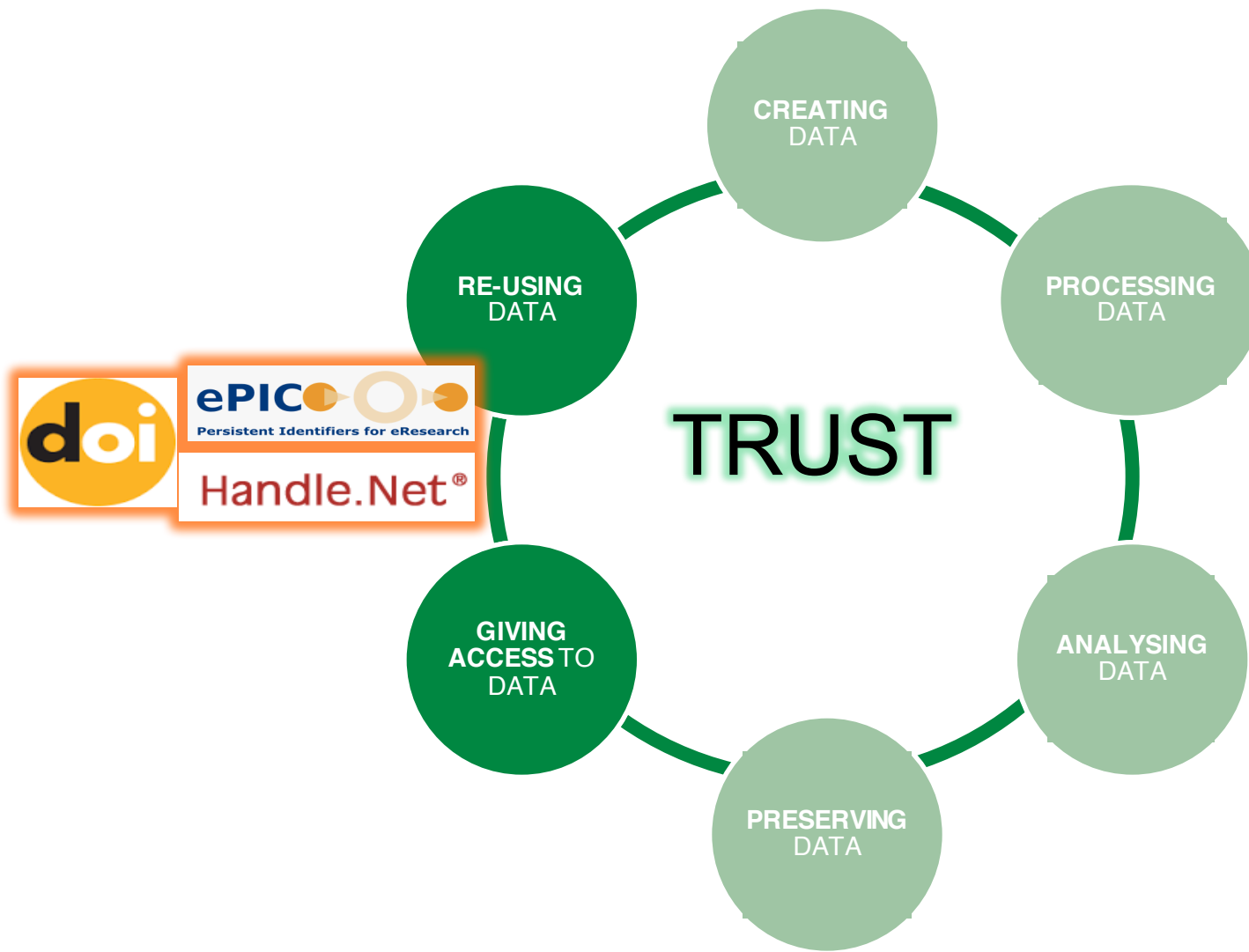
# SURFsara Data Archive

- 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
- Access: NWO grant, SURF e-infrastructure grant, or contract

# The data Life cycle






# PID Service

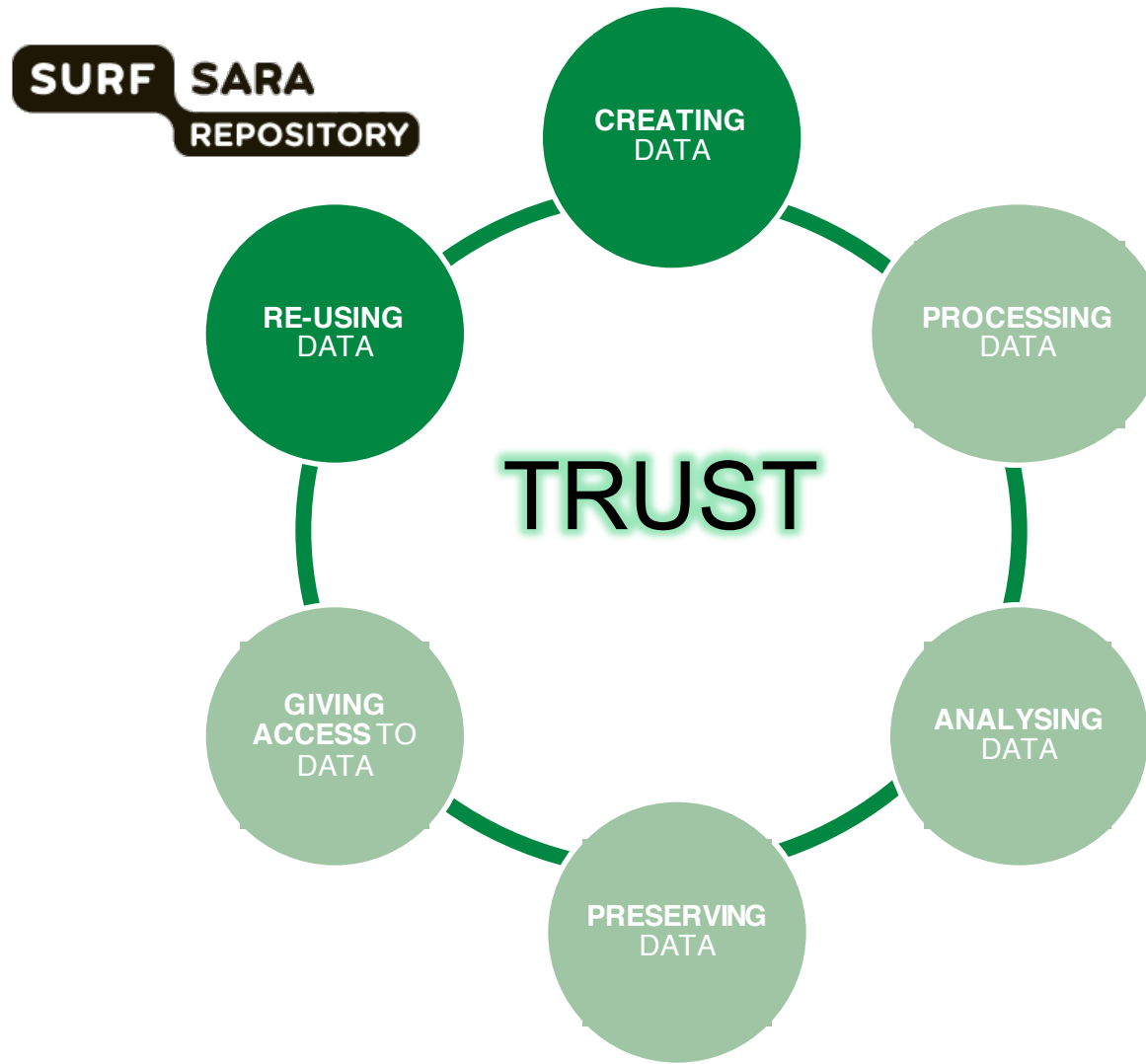
- PIDs (Persistent Identifiers) ensures the findability of your data
  - Pointers to resources like files, folders, webpages, real world objects
  - Globally unique
  - Resolvable via http
  - Comparable to ISBN numbers assigned to books
- Example resolvers: <https://dx.doi.org/> and <http://hdl.handle.net/>
- A PID consists of a prefix and a postfix (11304/2e873bd8-b988-11e3-8cd7-14feb57d12b9)



# PIDs – Handle, EPIC and DOIs

- Handle
  - Technology to create, store and update PIDs
  - Run by corporation of National Research Institutes (CNRI)
  - Infrastructure and technology to resolve PIDs
- EPIC (European Persistent Identifier Consortium)
  - Maintaining reliable PID service for storing data
  - Employing Handle technology
- DOI (Digital Object Identifier)
  - Based on Handle system
  - Well established in the publisher's world
- <http://www.ncdd.nl/pid-wijzer/>

# The data Life cycle



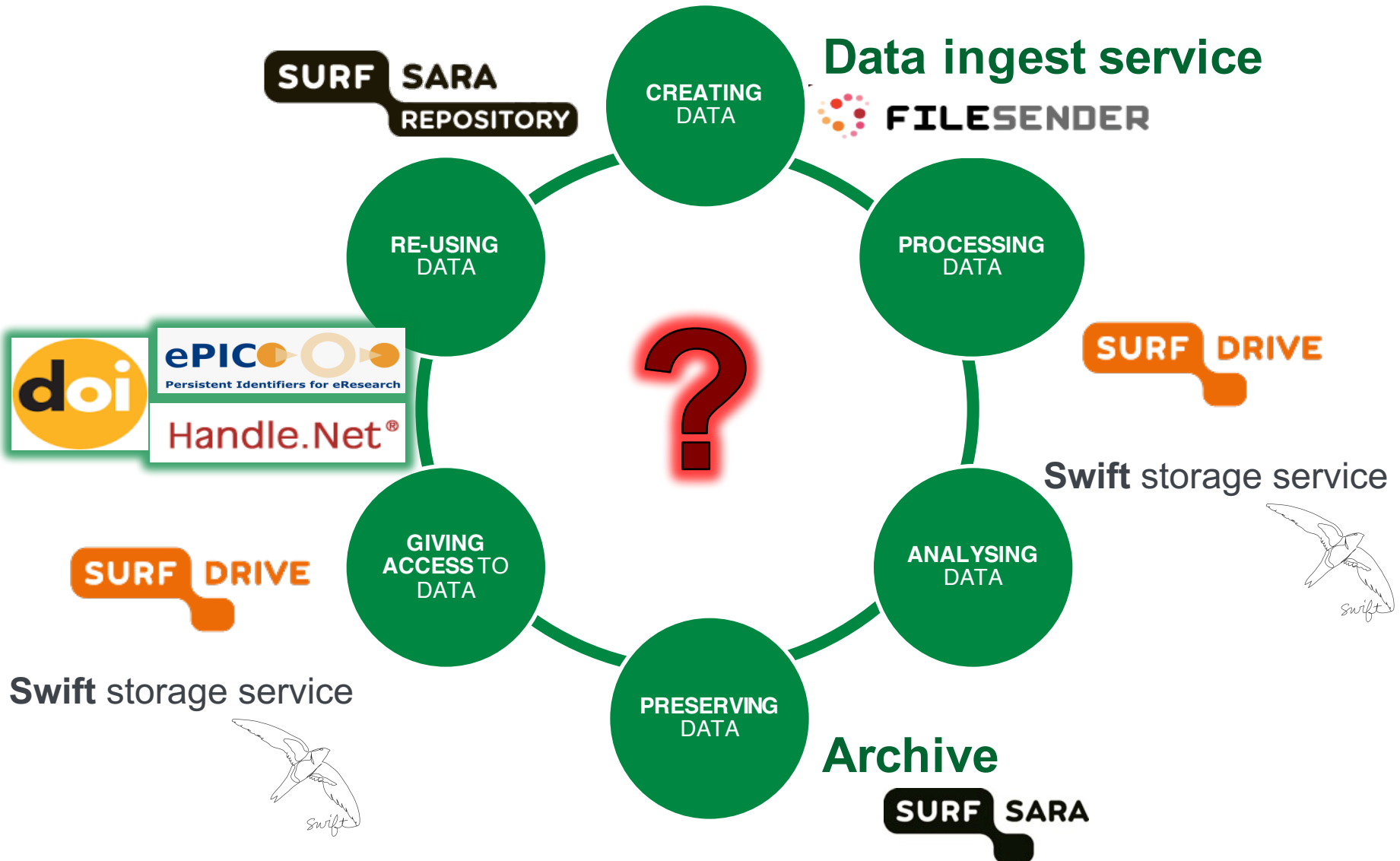
# SURFsara Data Repository

- Data repository service to deposit and publish data
- Long-term preservation of research data
- Provides quality to data sets and objects via metadata descriptions
- Makes data citable and findable via Persistent Identifiers
- Status: Under development





# The data Life cycle



# Thank You!



Thanks to:

Christine Staiger (SURFsara)

Hans van Piggelen (SURFsara)

Arthur Newton (SURFsara)