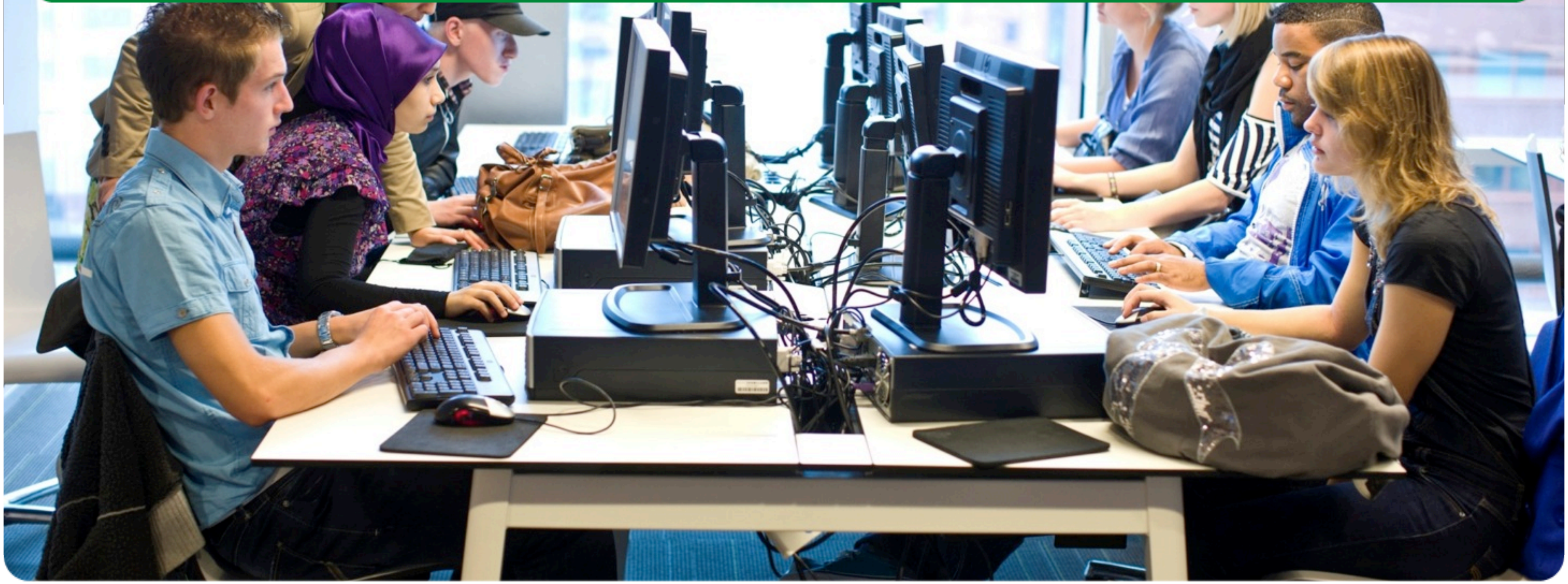
Data Management Services

NARGES ZARRABI



UvA HPC Course – 29th January 2018

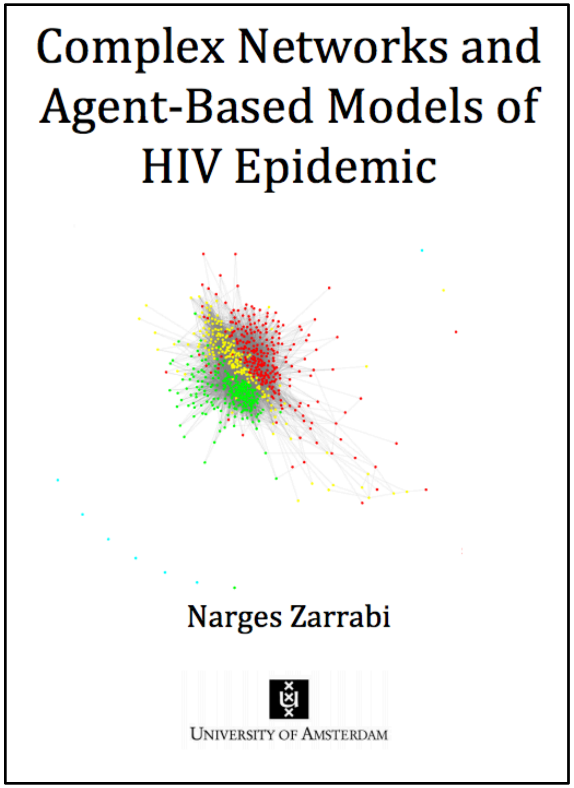
SURF SARA

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)

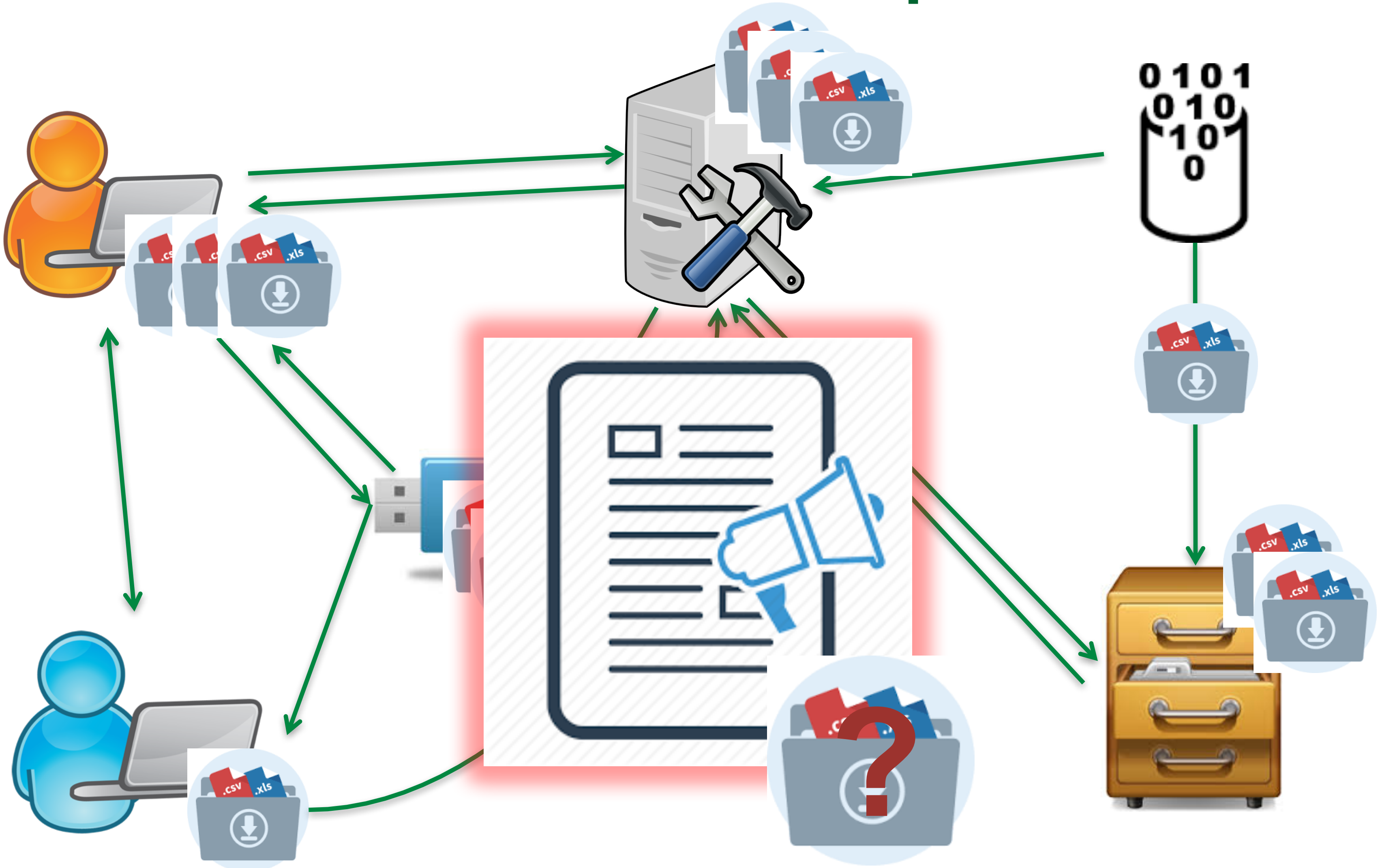


Complex Networks and Agent-Based Models of HIV Epidemic

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

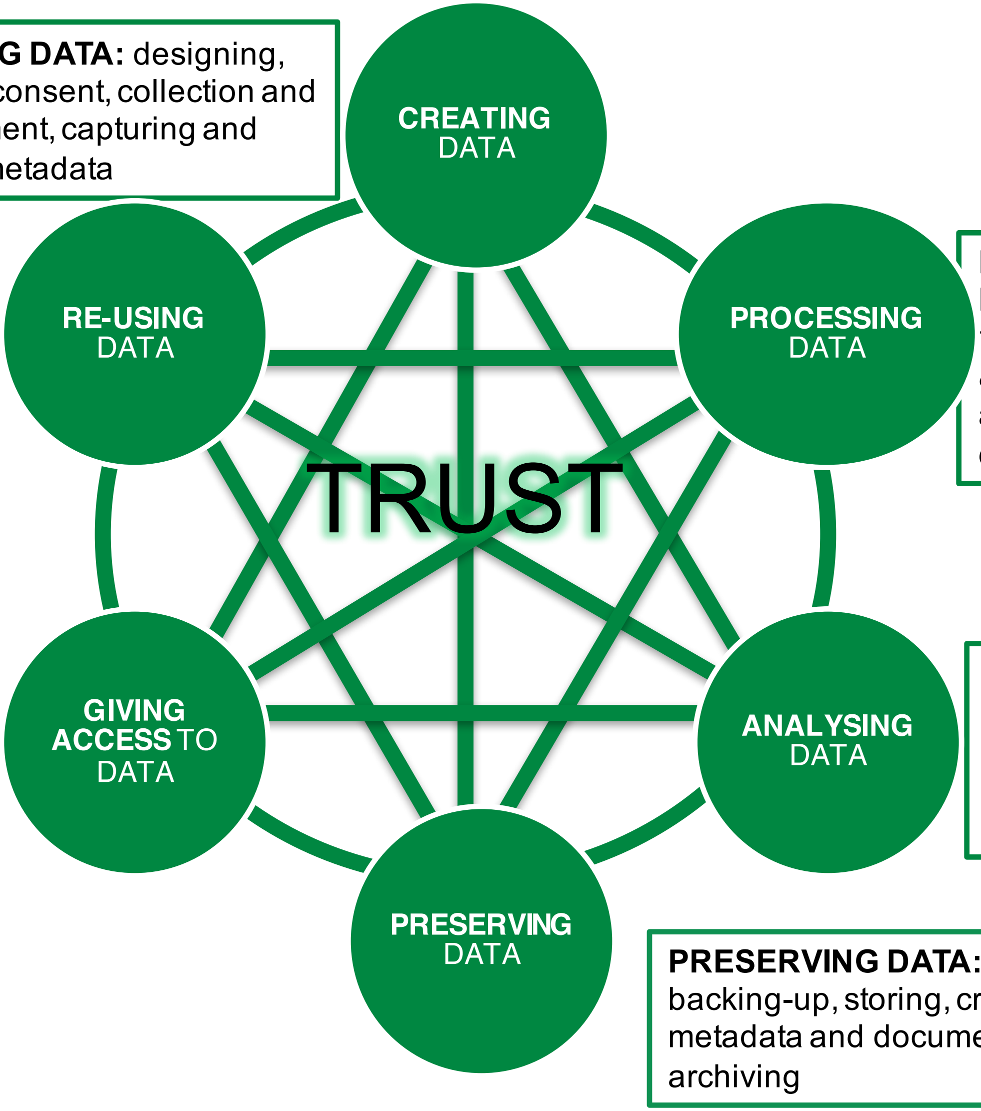
□ Type of data

□ Data & metadata standards

□ Sharing

□ Transition from collection to reuse

The data Life cycle



CREATING DATA

PROCESSING DATA

ANALYSING DATA

PRESERVING DATA

GIVING ACCESS TO DATA

RE-USING DATA

TRUST

CREATING DATA: designing, planning consent, collection and management, capturing and creating metadata

PROCESSING DATA: entering, transcribing, checking & validating, anonymizing and describing

ANALYSING DATA: interpreting, deriving, producing outputs & publishing, preparing for sharing

PRESERVING DATA: migrating, backing-up, storing, creating metadata and documentation, archiving

ACCESS TO DATA: distributing, sharing, controlling access, promoting

RE-USING DATA: for follow-ups, new research, research reviews, scrutinising, teaching & learning

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

ePIC
Persistent Identifiers for eResearch

Data Archiving and Networked Services
DANS
EASY

SURF DRIVE

4TU.Centre for Research Data

FILESENDER

iRODS

doi

YODA

B2SAFE

The Dataverse Project

Handle.Net®

SWIFT storage service

swift

Archive
SURF SARA

Data Archiving and Networked Services
DANS
NARCIS

Data ingest service

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

ePIC
Persistent Identifiers for eResearch

SURF DRIVE

Handle.Net®

FILESENDER

SWIFT storage service

swift

Archive

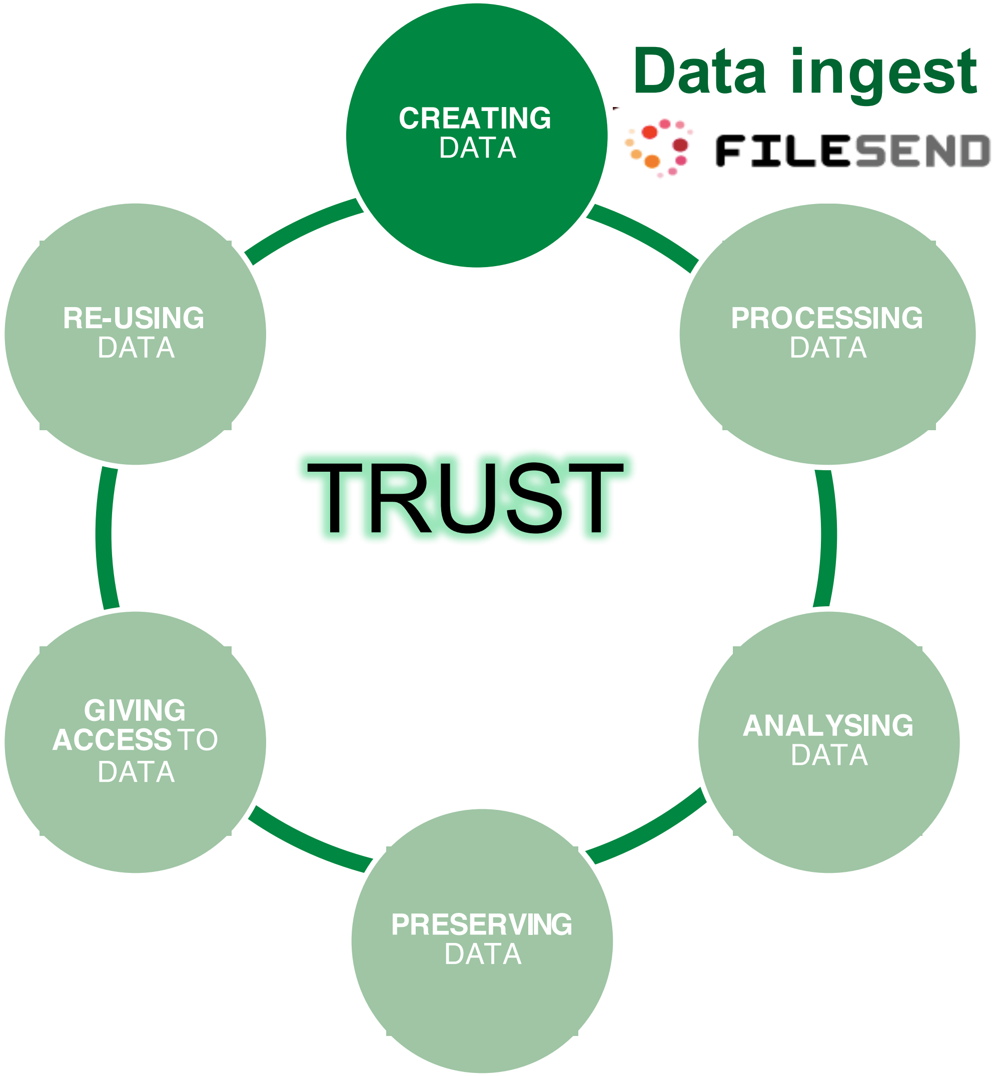
SURF SARA

Data ingest service

The data life cycle

Data ingest service

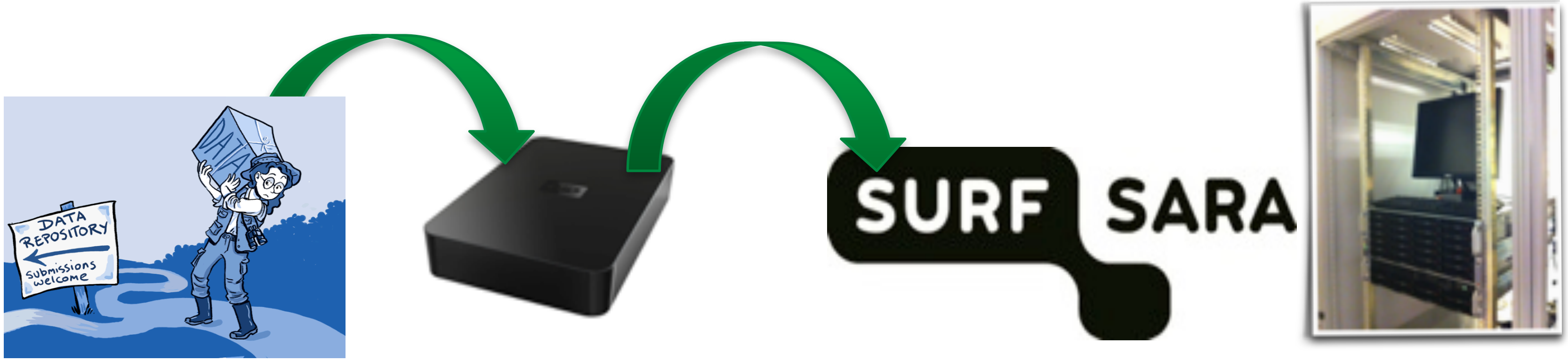
FILESENDER



Data ingest service

- Data often resides on external storage media, USB sticks, external hard drives
- Slow or no internet connection

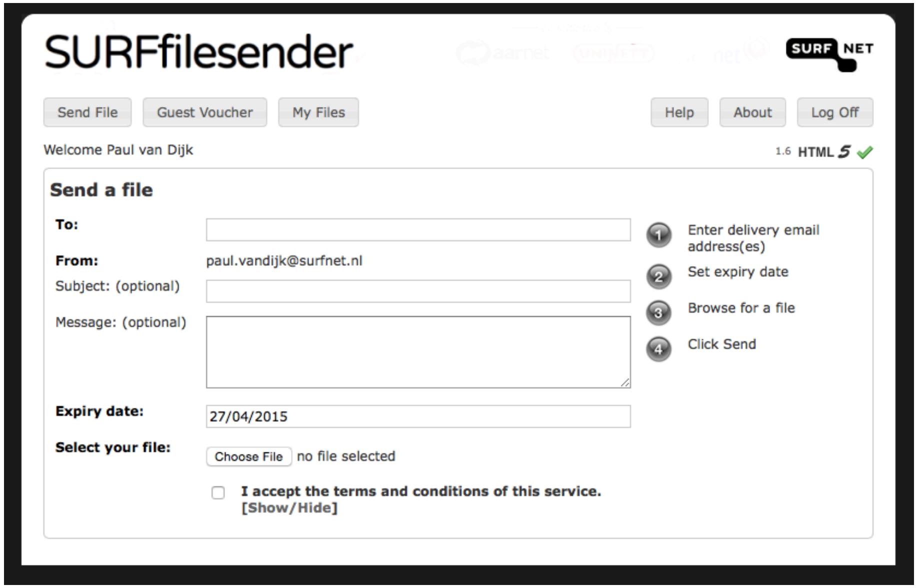
- Easy way to upload large data from disk to SURFsara facilities
- Upload data from 45 disks in parallel



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/



SURFfilesender

Welcome Paul van Dijk

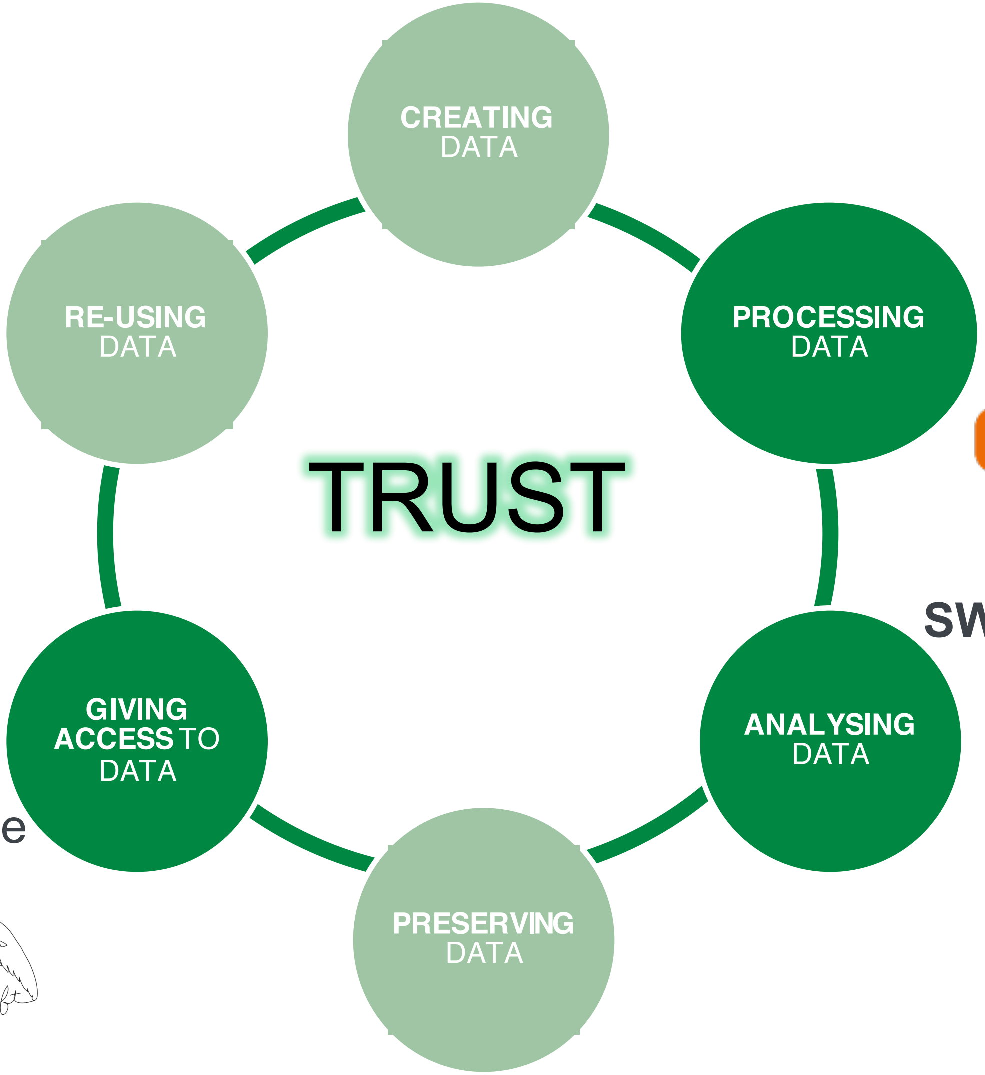
Send a file

To:	
From:	paul.vandijk@surfnet.nl
Subject: (optional)	
Message: (optional)	
Expiry date:	27/04/2015
Select your file:	Choose File no file selected

□ **I accept the terms and conditions of this service. [Show/Hide]**

1. Enter delivery email address(es)
2. Set expiry date
3. Browse for a file
4. Click Send

The data Life cycle



FILESENDER

SURF DRIVE

SWIFT storage service

SURF DRIVE

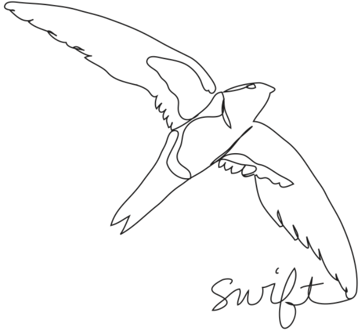
SWIFT storage service

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

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
- Documentation: https://doc.swift.surfsara.nl/en/latest/

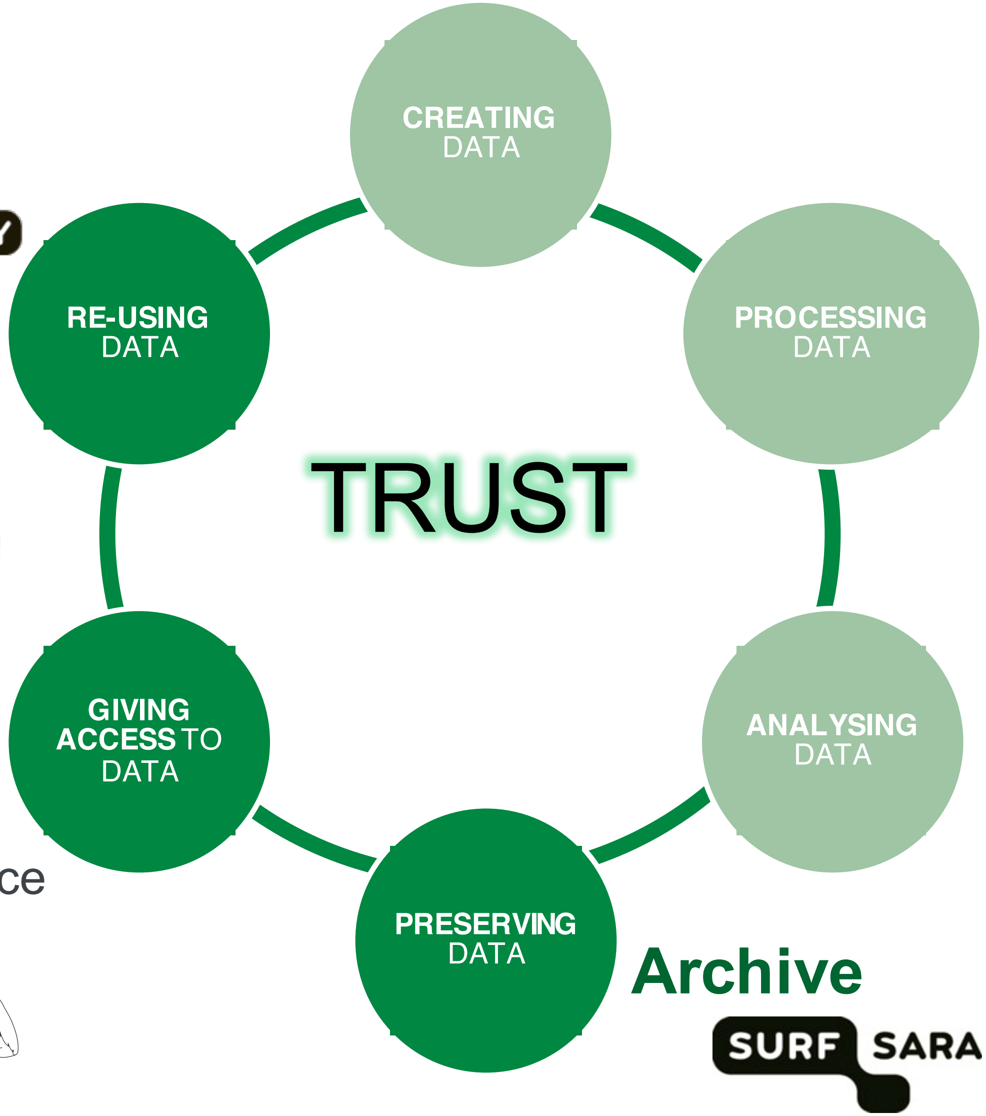
The data Life cycle

SURF SARA REPOSITORY

FILESENDER

SURF DRIVE

SWIFT storage service



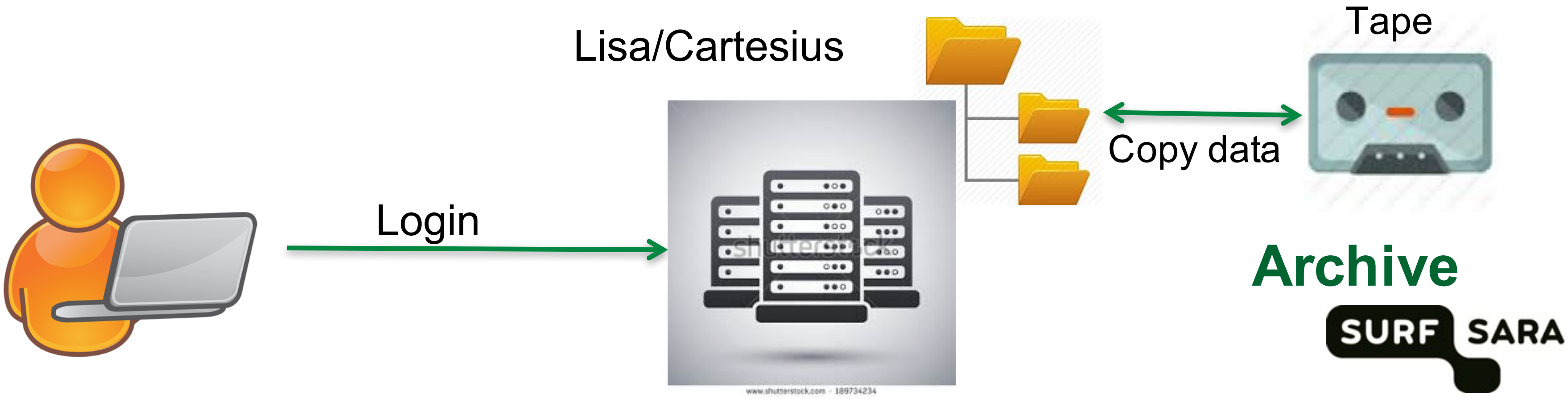
Archive

SURF SARA

SURFsara Data Archive

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

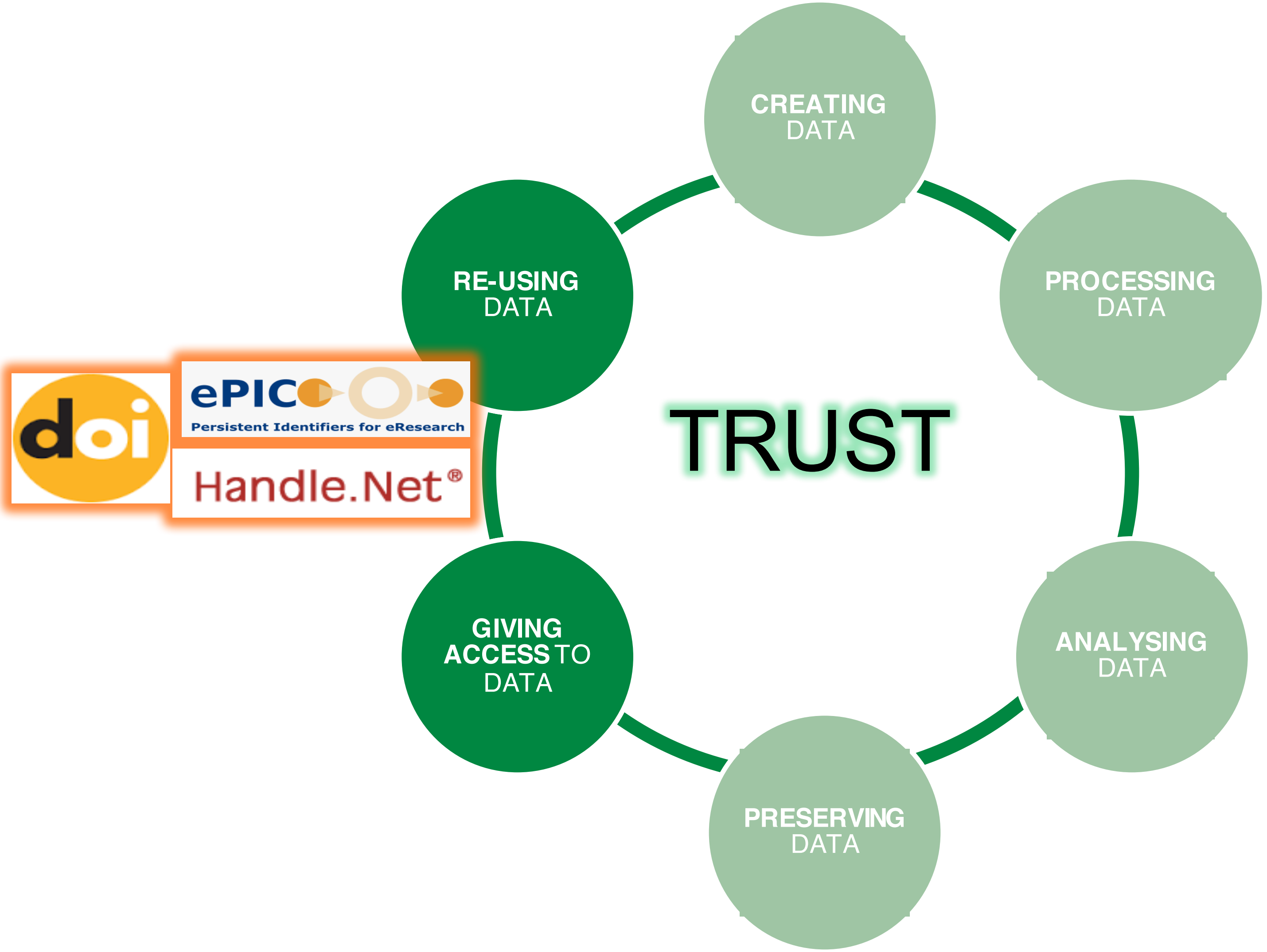
- Powerful transfer protocols:
  - gridfTP
  - rsync
  - scp



- 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

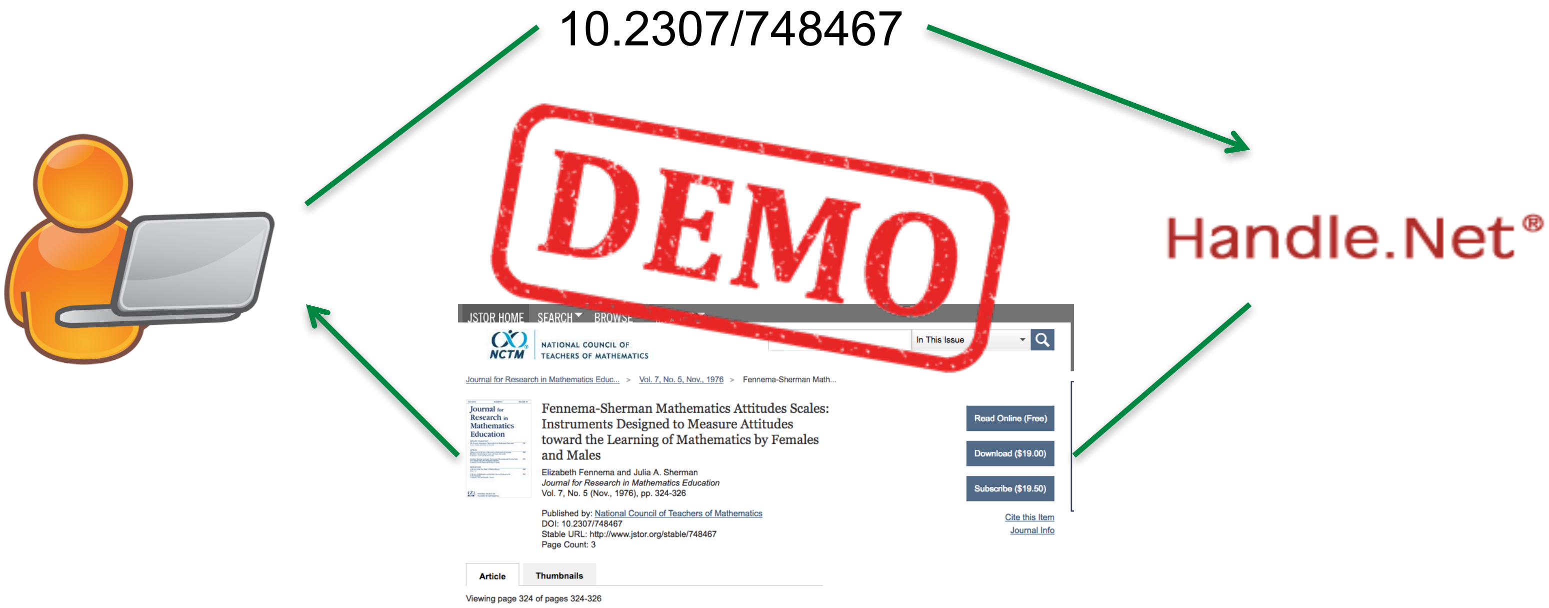
The data Life cycle



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)



Journal for Research in Mathematics Educ... > Vol. 7, No. 5, Nov., 1976 > Fennema-Sherman Math...

Fennema-Sherman Mathematics Attitudes Scales: Instruments Designed to Measure Attitudes toward the Learning of Mathematics by Females and Males

Elizabeth Fennema and Julia A. Sherman
Journal for Research in Mathematics Education
Vol. 7, No. 5 (Nov., 1976), pp. 324-326

Published by: National Council of Teachers of Mathematics
DOI: 10.2307/748467
Stable URL: http://www.jstor.org/stable/748467
Page Count: 3

Viewing page 324 of pages 324-326

PIDs – Handle, EPIC and DOIs

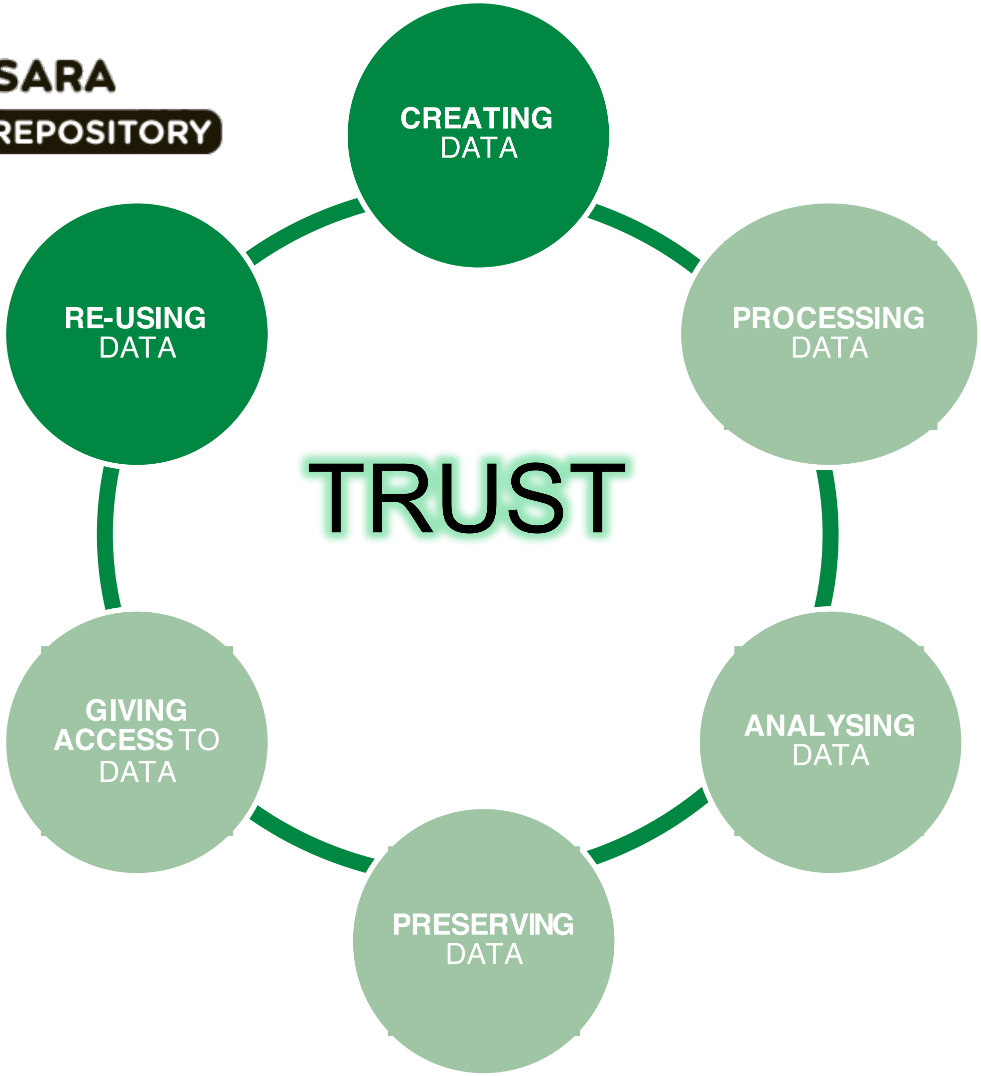
- Handle
  - Technology to create, store and update PIDs
  - Run by corporation of National Research Institutes (CNRI)
  - Infrastructure and technology to resolve PIDs

- 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 pulisher's world

- **http://www.ncdd.nl/pid-wijzer/**

The data Life cycle



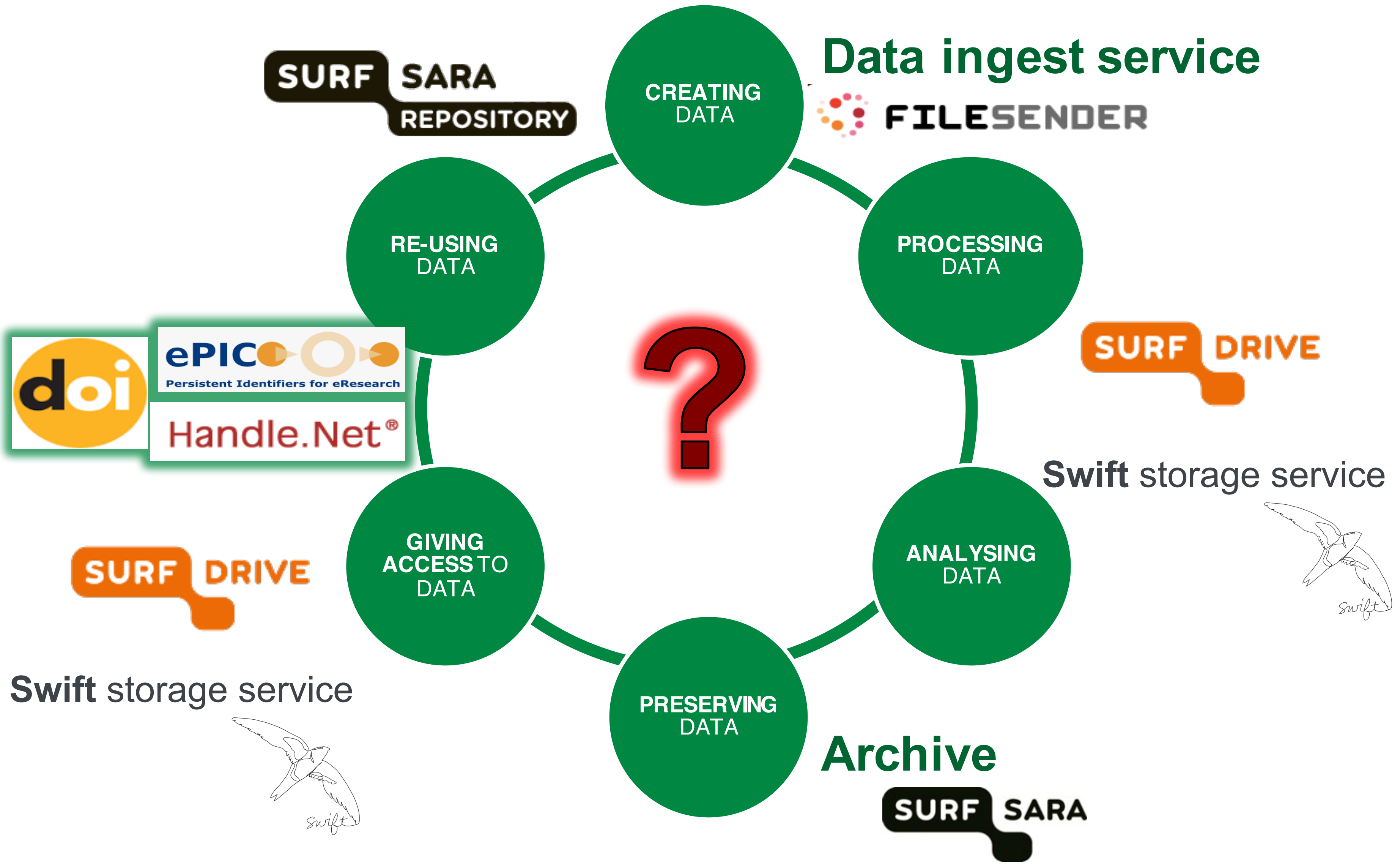
SURFsara Data Repository

- 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

dataseal ofapproval 2010

SURF SARA REPOSITORY

The data Life cycle



CREATING DATA

Data ingest service

PROCESSING DATA

ANALYSING DATA

Swift storage service

PRESERVING DATA

Archive

GIVING ACCESS TO DATA

Swift storage service

RE-USING DATA

?

Thank You!

Thanks to:
Christine Staiger (SURFsara)