Data services

Christine Staiger, SURFsara

Agenda

NEW

1pm - 2.10pm

- Introduction (Christine)
- GUI to archive, Cyberduck and FileZilla (Narges)
- **Coffee break**

2.10pm - 4.10pm

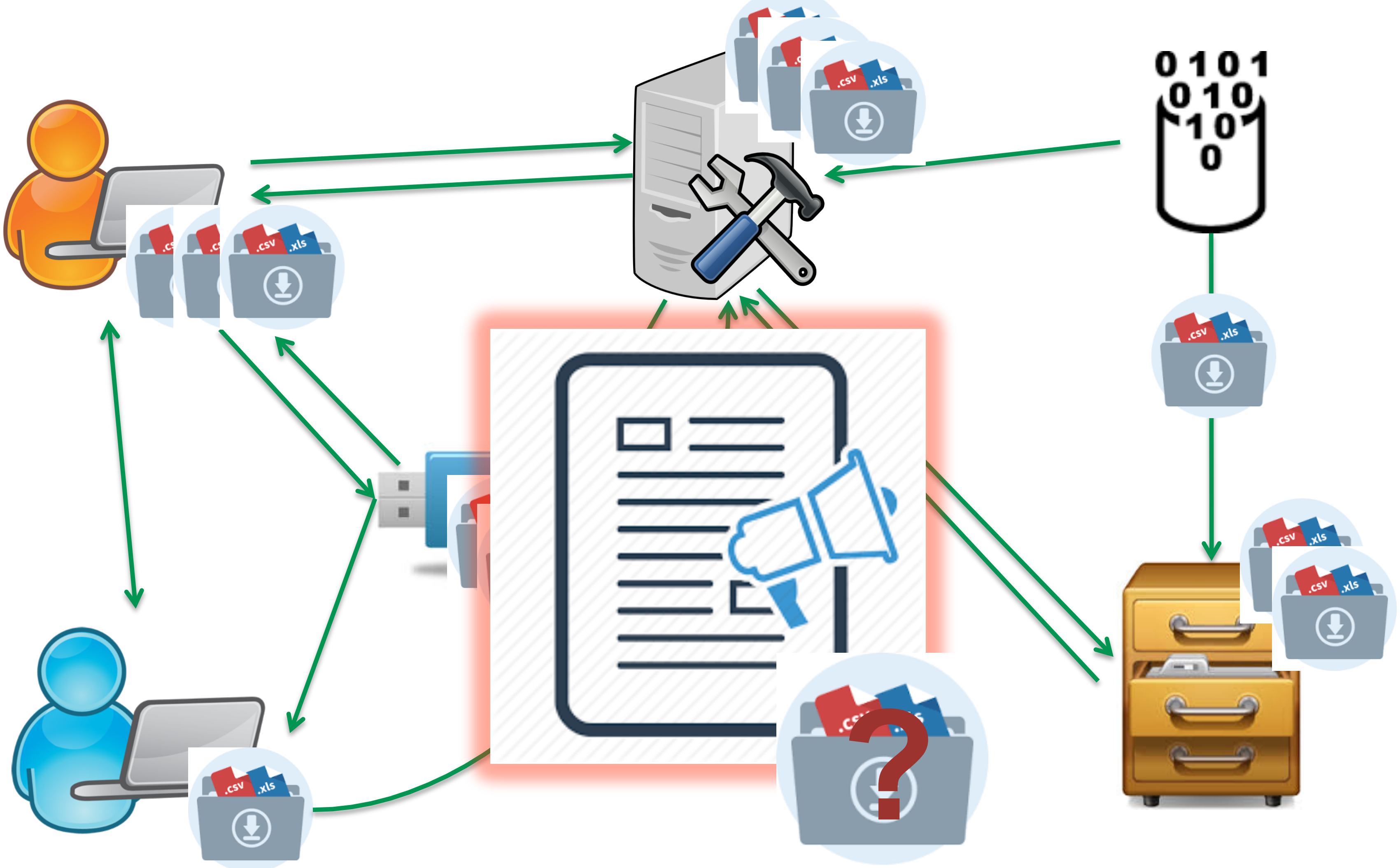
- scp + rsync (Jeroen)
- Tarring data with tar and dmftar (Hans)
- **Coffee break**

4.10pm - 5pm

- Wrap-up (Christine)
- Final Challenge:

  Putting it all together in a python workflow

Data – where is the problem?



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

FILESENDER

4TU.Centre for Research Data

doi

yoda

B2FIND

The Dataverse Project

Handle.Net®

Data ingest service

Archive

SURF SARA

Data Archiving and Networked Services

DANS

NARCIS

BeeHub

Data services at SURFsara

ePIC
Persistent Identifiers for eResearch

SURF DRIVE

FILESENDER

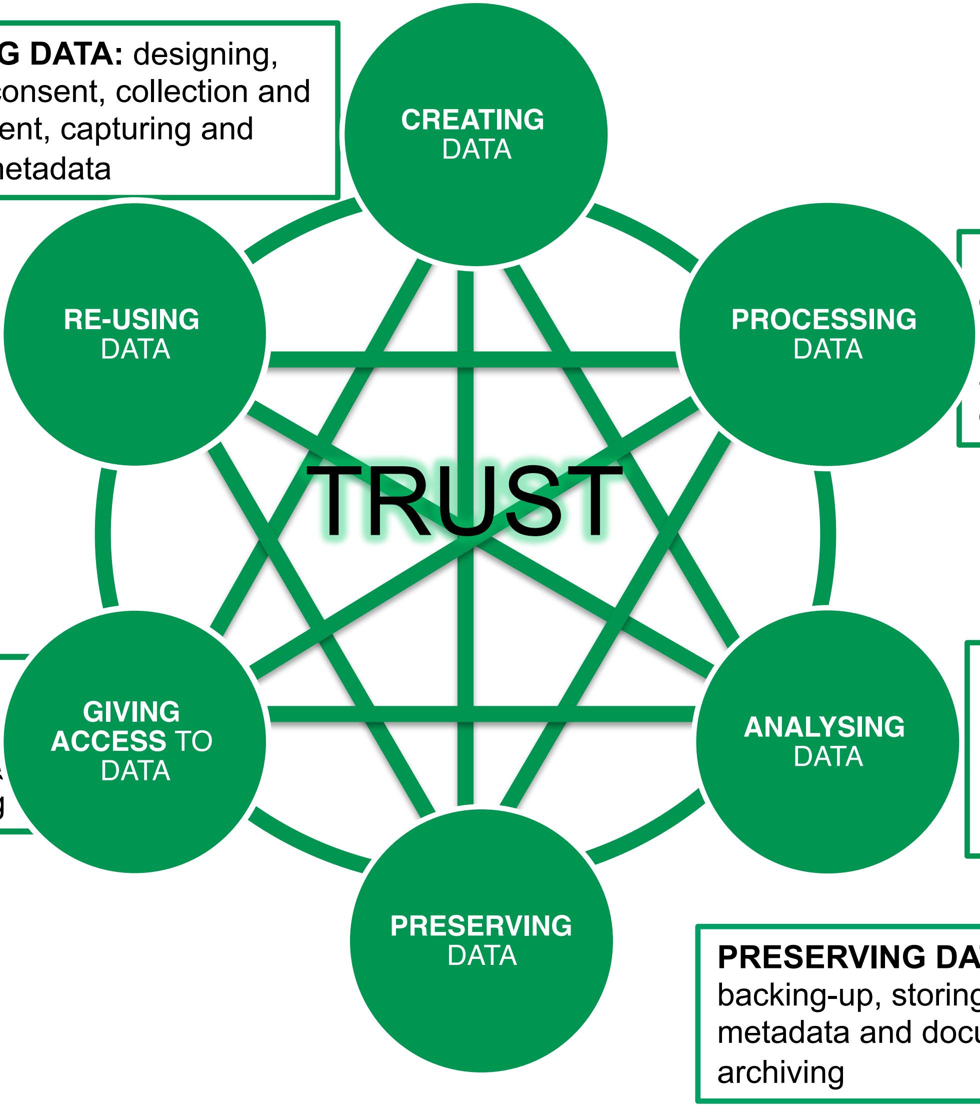
Data ingest service

Handle.Net®

Archive
SURF SARA

BeeHub

The data Life cycle



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

PROCESSING DATA: entering, transcribing, checking & validating, anonymising 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

The researchers’ needs

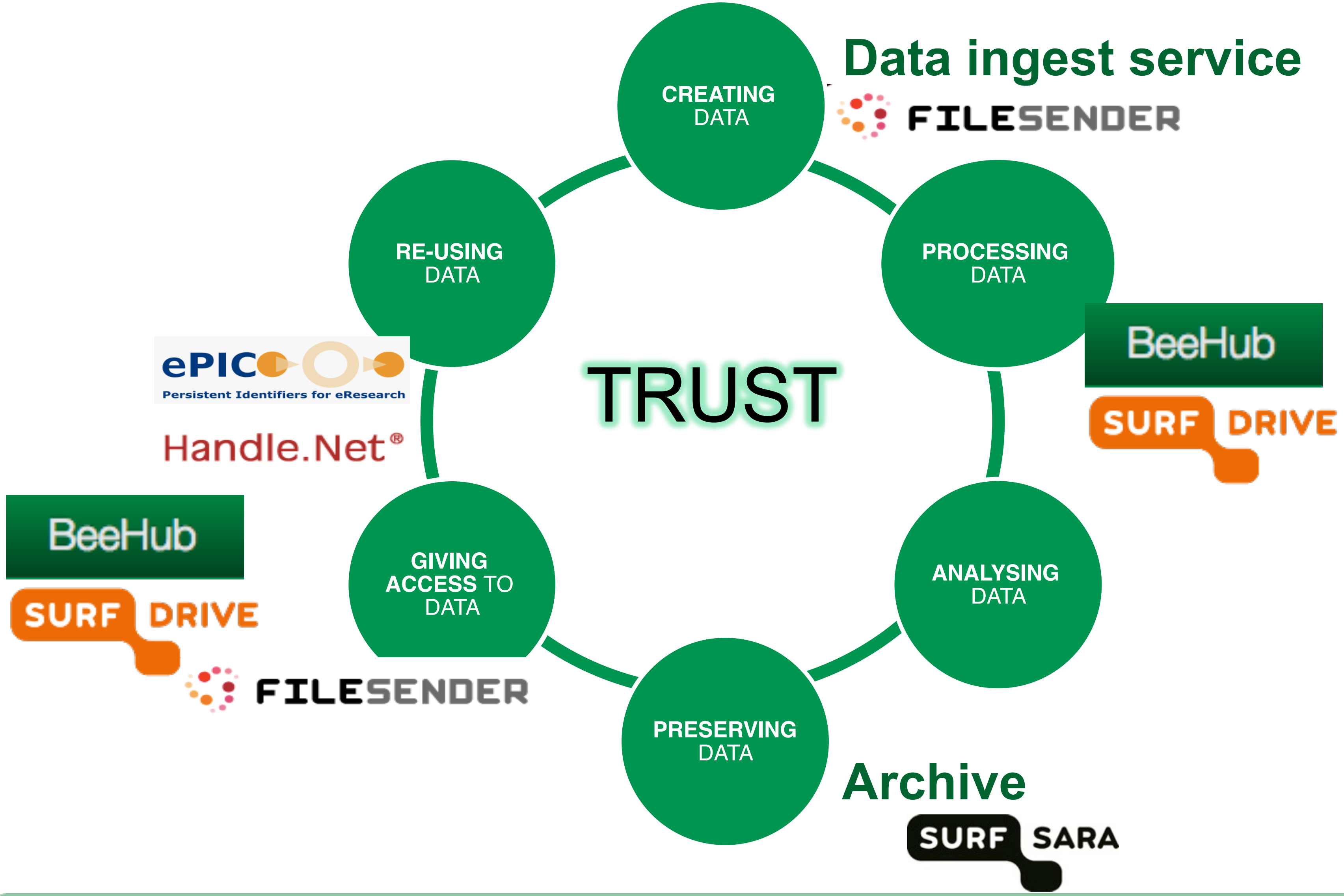
- **Store** data during research
- **Share** data during and after research
- **Archive** data
- **Synchronise** data across different locations

- **Link** publication to processed and raw data
- **Publish** data
- **Find** data and **make data findable** by others

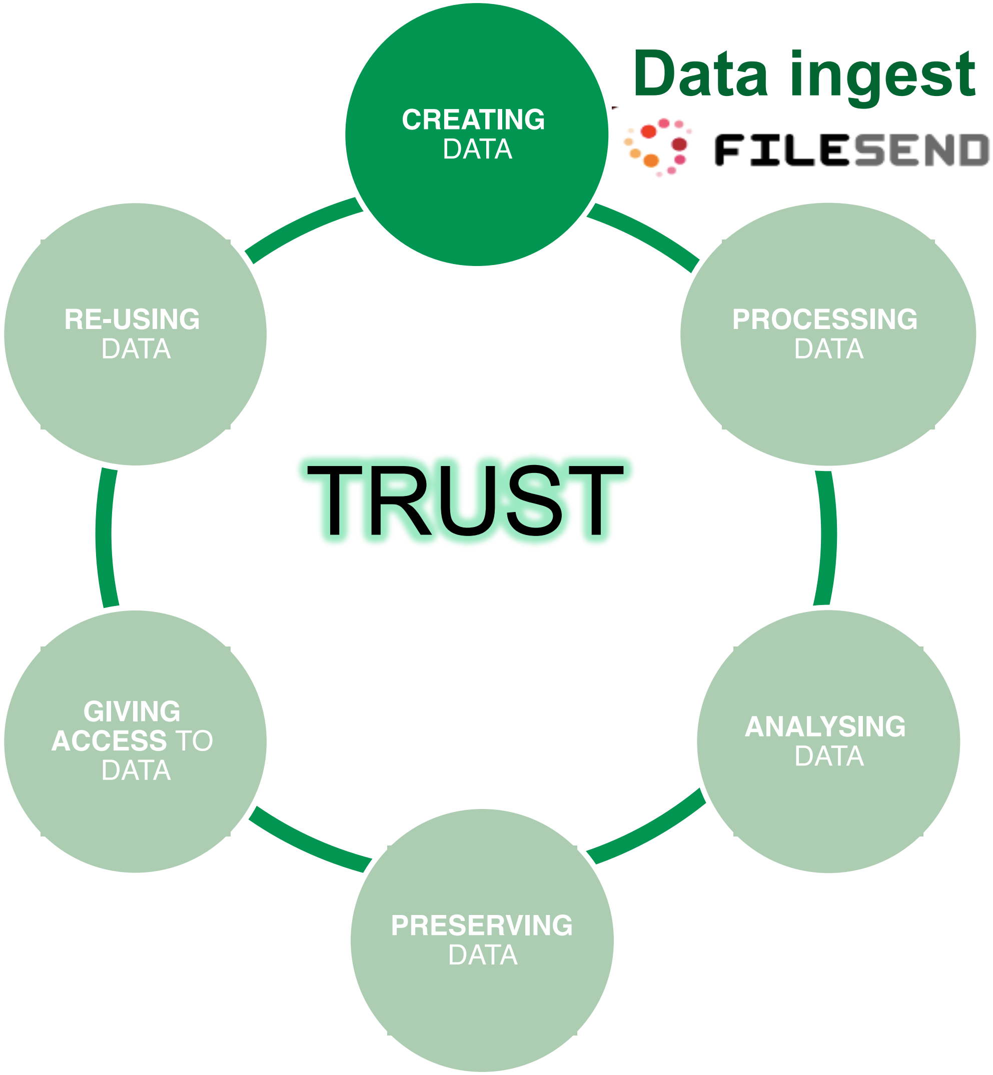
- Data **transfers**
- Data **provenance**: what happened with the data

- …

The data Life cycle



The data Life cycle



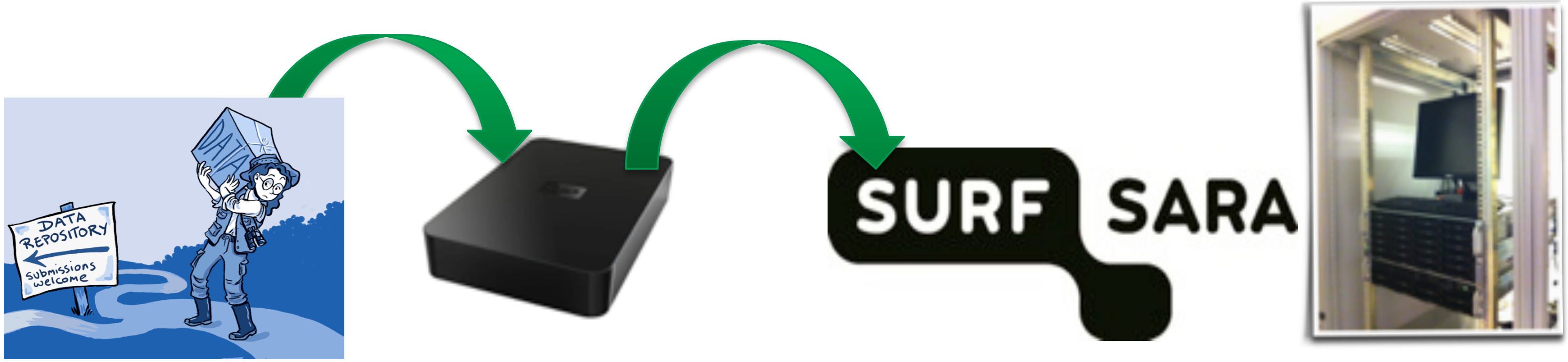
Data ingest service

FILESENDER

Data ingest service

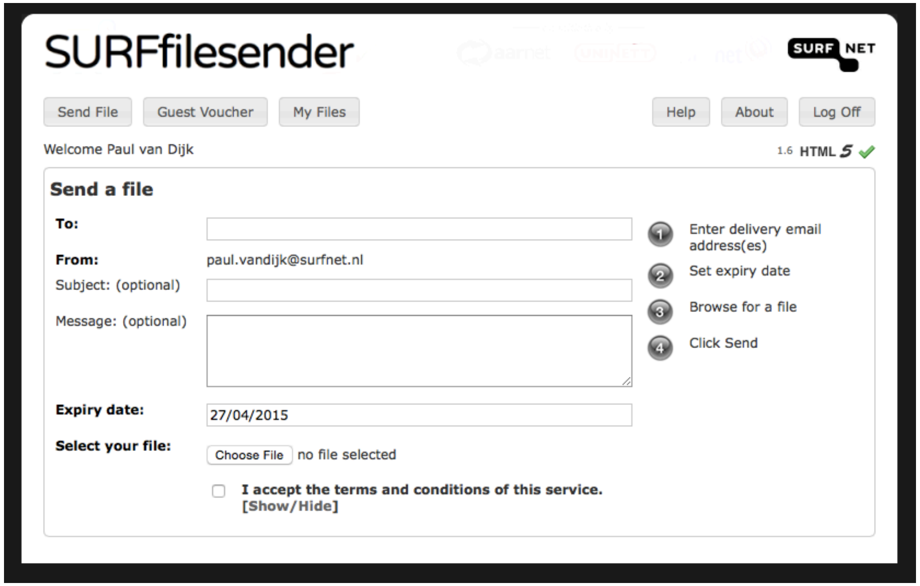
- Data often resides on external storage media, USB sticks, external hard drives
- Slow or no internet connection

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



FILESENDER

- Trusted community service
- Transferring BIG files from person to person
- File Transport service not File Storage (!)
- Simple interface
- Up to 200GB per file
- Option to apply end-to-end encryption (250MB browser limit)
- Vouchers for guest usage



Welcome Paul van Dijk

Send a file

To:

From: paul.vandijk@surfnet.nl

Subject: (optional)

Message: (optional)

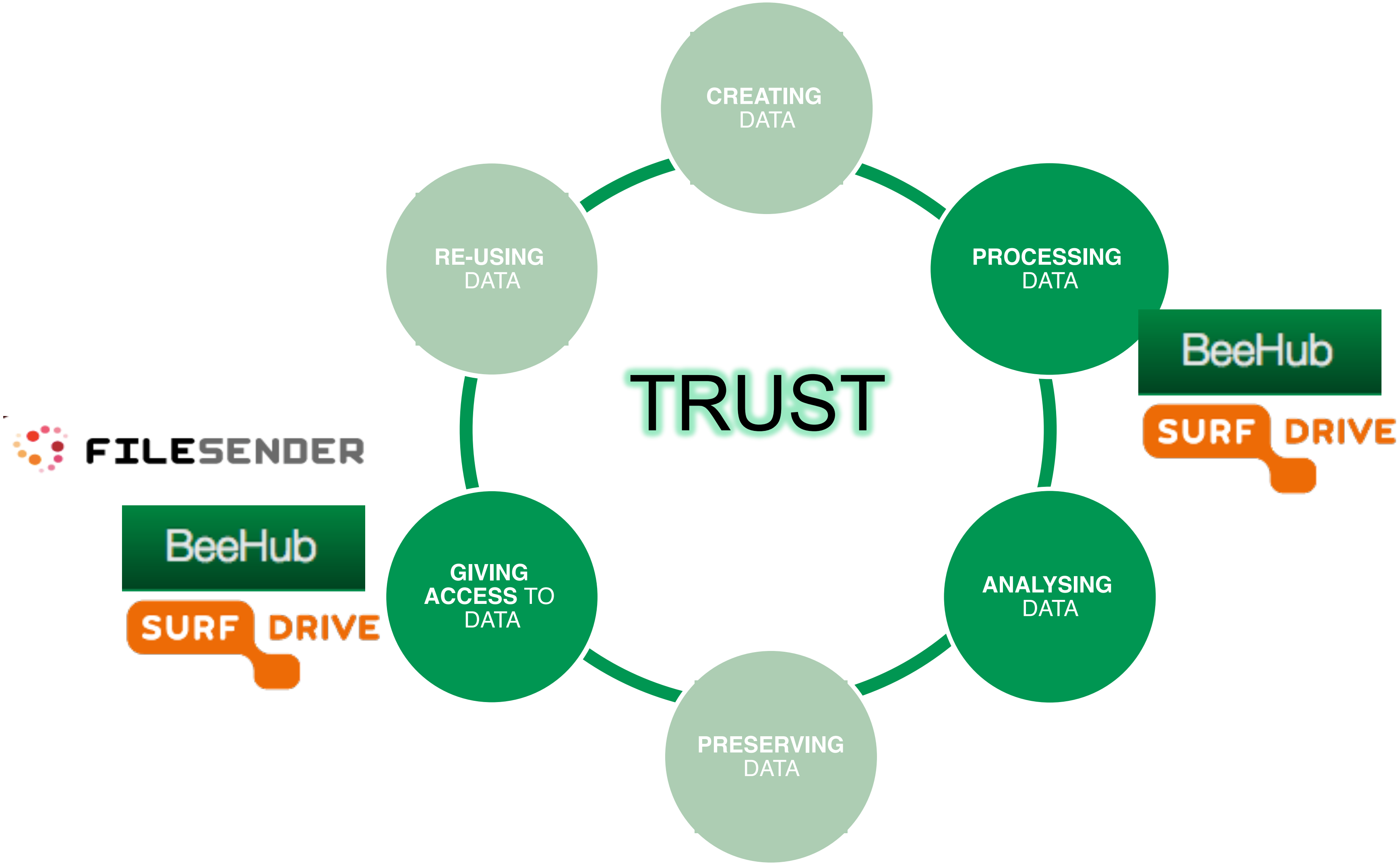
Expiry date: 27/04/2015

Select your file: Choose File no file selected

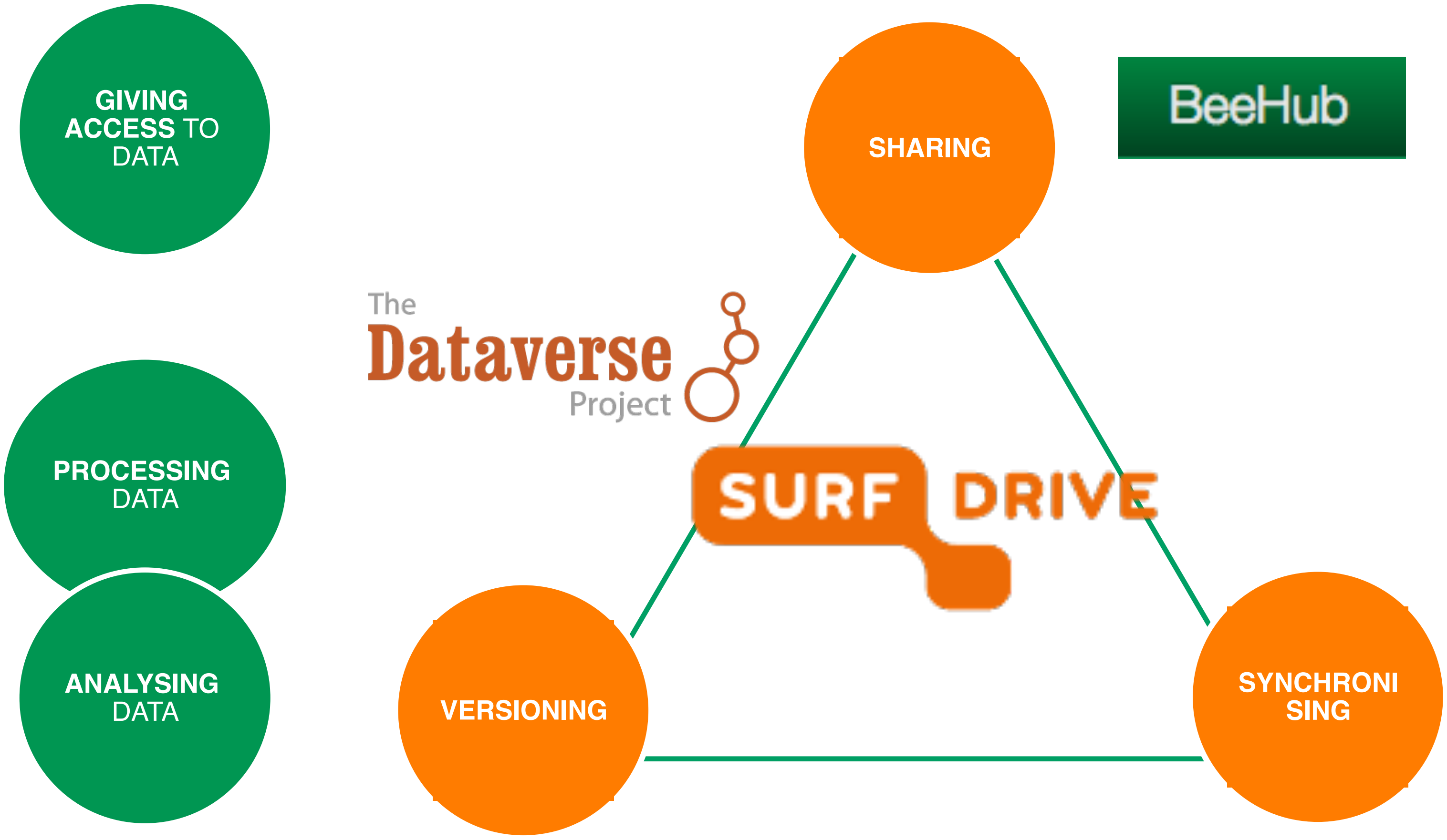
□ **I accept the terms and conditions of this service. [Show/Hide]**

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

The data Life cycle



Data services – what goes where?



GIVING ACCESS TO DATA

PROCESSING DATA

ANALYSING DATA

SHARING

VERSIONING

SYNCHRONISING

SURFdrive

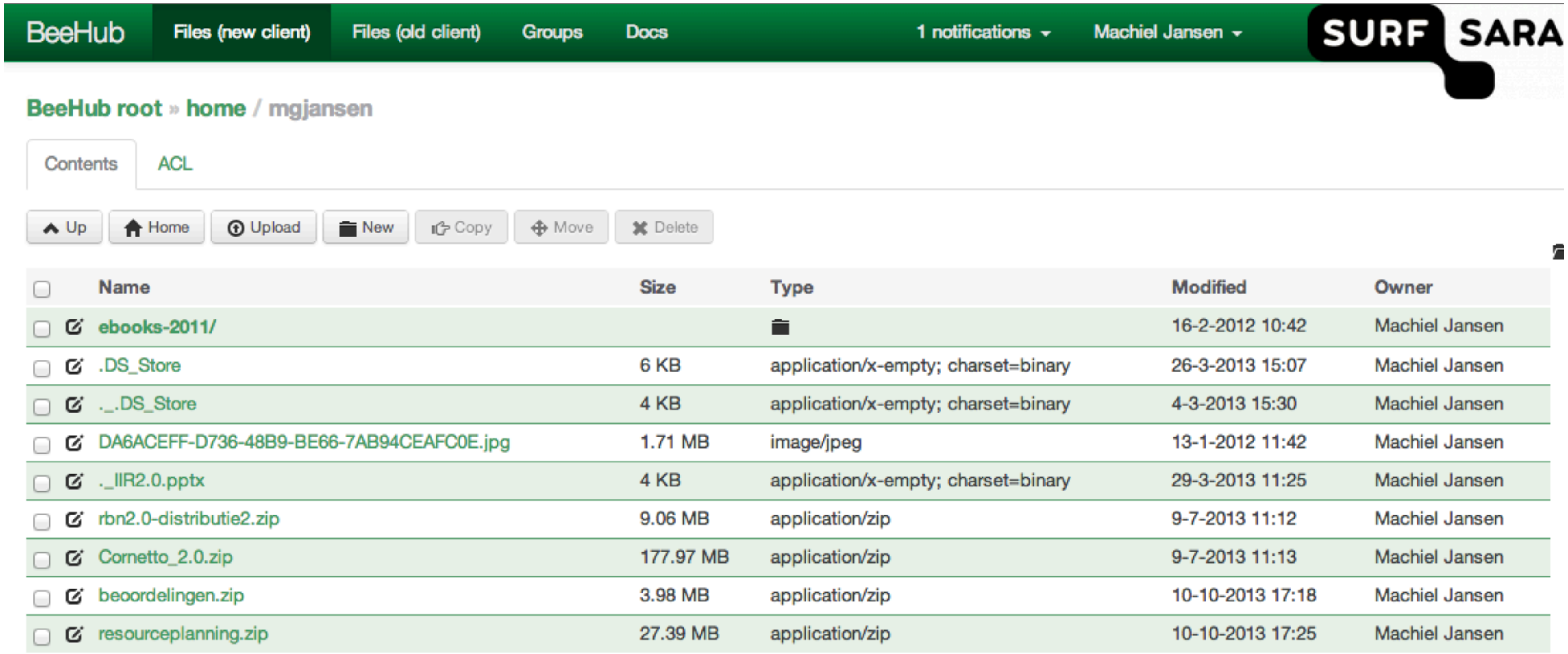
- Trusted community cloud for personal storage
- Sharing smaller data files

- Collaboration between SURFsara, SURFnet and Dutch universities
- Specifications and service determined by end-users (universities)

- 100 GB storage capacity per user
- Based on ownCloud, synchronises with local storage

SURF DRIVE

Beehub



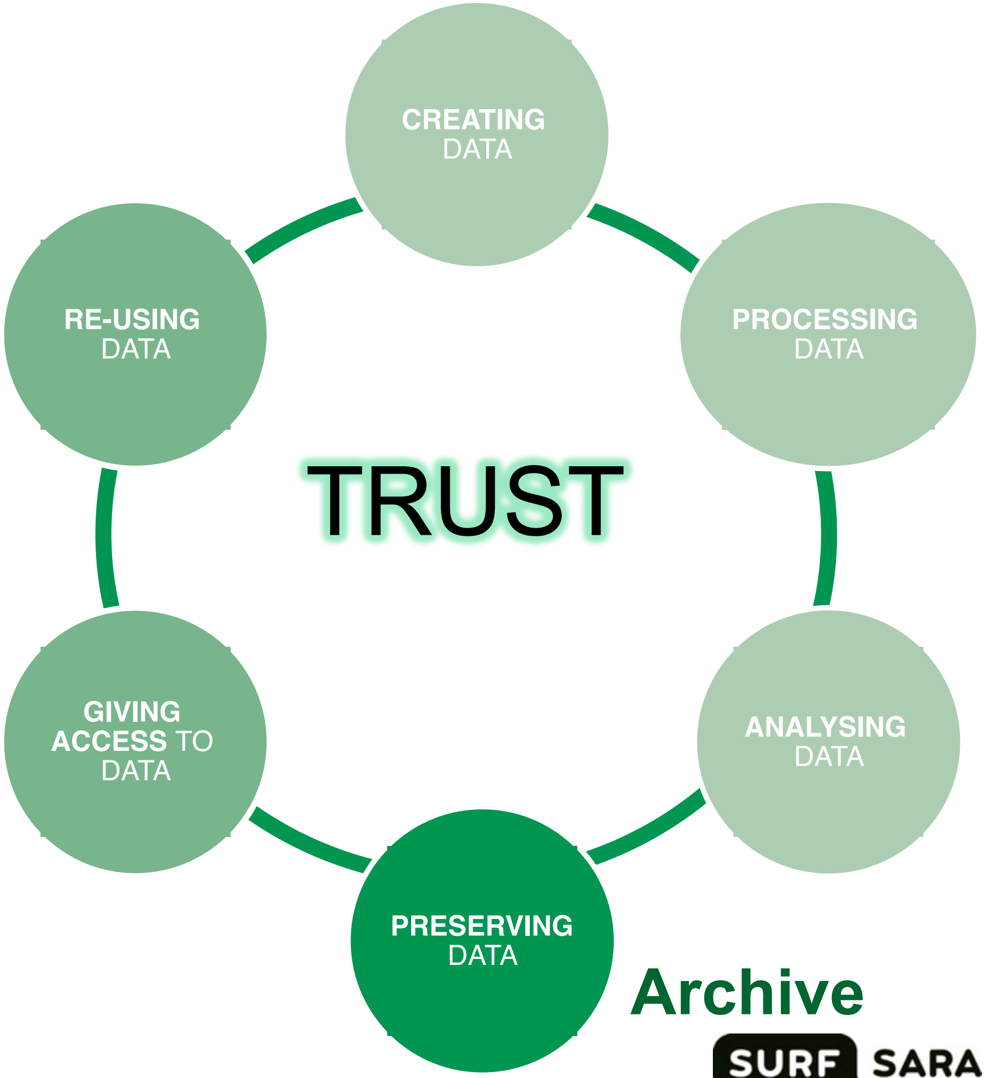
BeeHub root » home / mgjansen

□	Name	Size	Type	Modified	Owner
□	ebooks-2011/			16-2-2012 10:42	Machiel Jansen
□	.DS_Store	6 KB	application/x-empty; charset=binary	26-3-2013 15:07	Machiel Jansen
□	._.DS_Store	4 KB	application/x-empty; charset=binary	4-3-2013 15:30	Machiel Jansen
□	DA6ACEFF-D736-48B9-BE66-7AB94CEAFC0E.jpg	1.71 MB	image/jpeg	13-1-2012 11:42	Machiel Jansen
□	._IIR2.0.pptx	4 KB	application/x-empty; charset=binary	29-3-2013 11:25	Machiel Jansen
□	rbn2.0-distributie2.zip	9.06 MB	application/zip	9-7-2013 11:12	Machiel Jansen
□	Cornetto_2.0.zip	177.97 MB	application/zip	9-7-2013 11:13	Machiel Jansen
□	beoordelingen.zip	3.98 MB	application/zip	10-10-2013 17:18	Machiel Jansen
□	resourceplanning.zip	27.39 MB	application/zip	10-10-2013 17:25	Machiel Jansen

- Sharing large data
- Mountable via webdav (no synchronisation)
- Capacity: 80 TB

- Temporary storage
- 100GB/user
- You can apply for more via SURF e-infrastructure grant

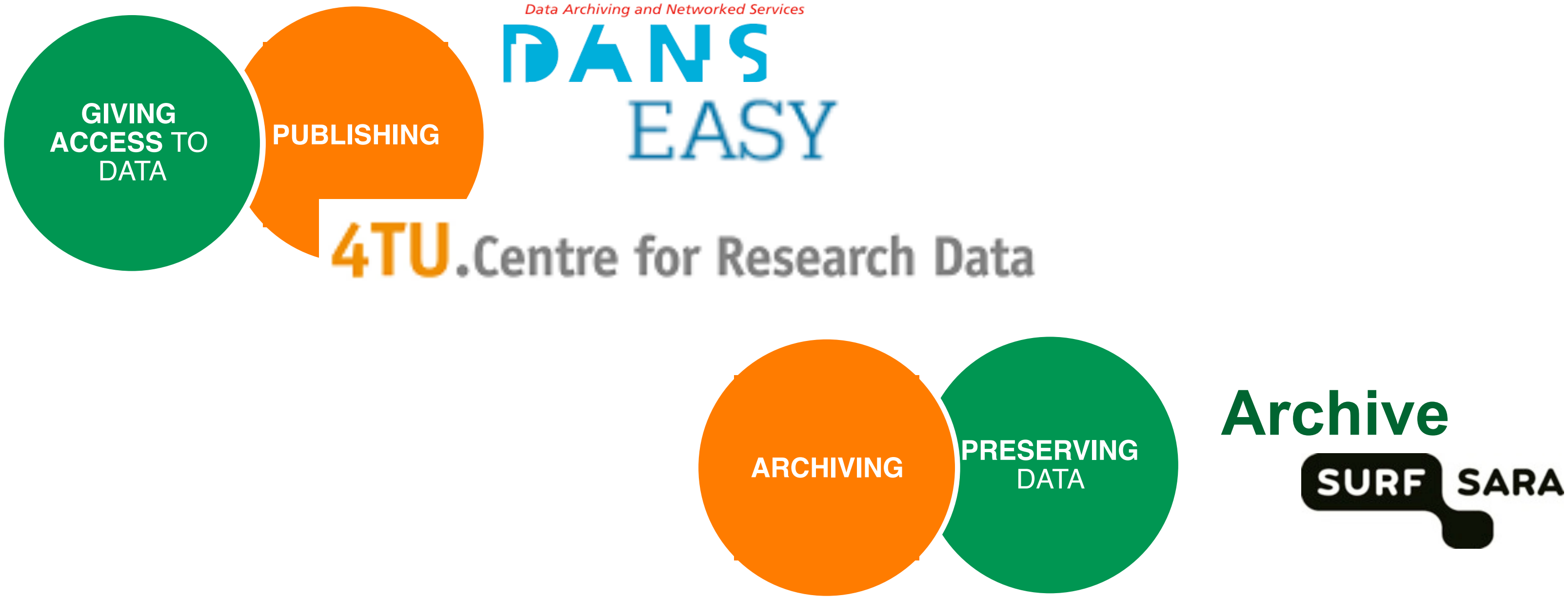
The data Life cycle



Archive

SURF SARA

Data services – what goes where?



GIVING ACCESS TO DATA

PUBLISHING

Data Archiving and Networked Services

DANS

EASY

4TU.Centre for Research Data

ARCHIVING

PRESERVING DATA

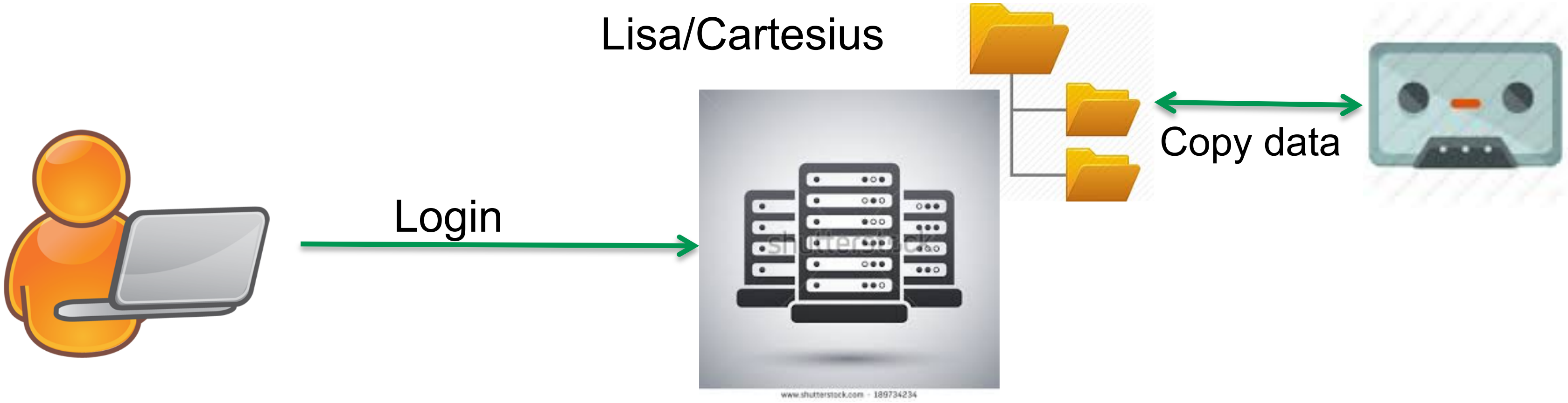
Archive

SURF SARA

The Archive – Long-term storage

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

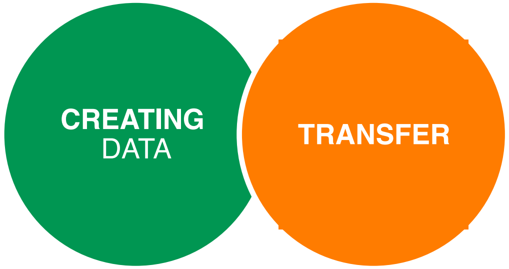
- Powerful transfer protocols:
  - gridfTP
  - rsync
  - scp



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

- Access: NWO grant or SURF e-infrastructure grant

Data services – what goes where?



Data ingest service

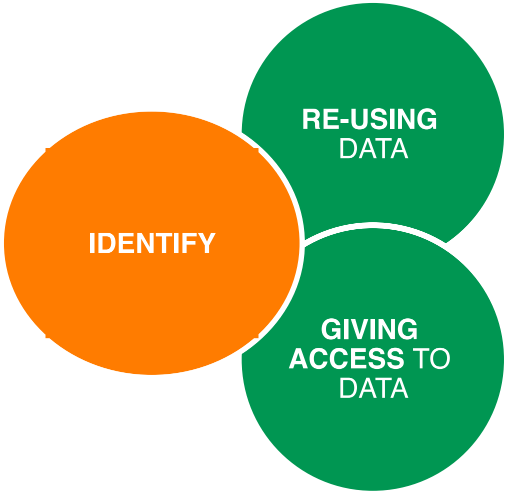
FILESENDER

doi

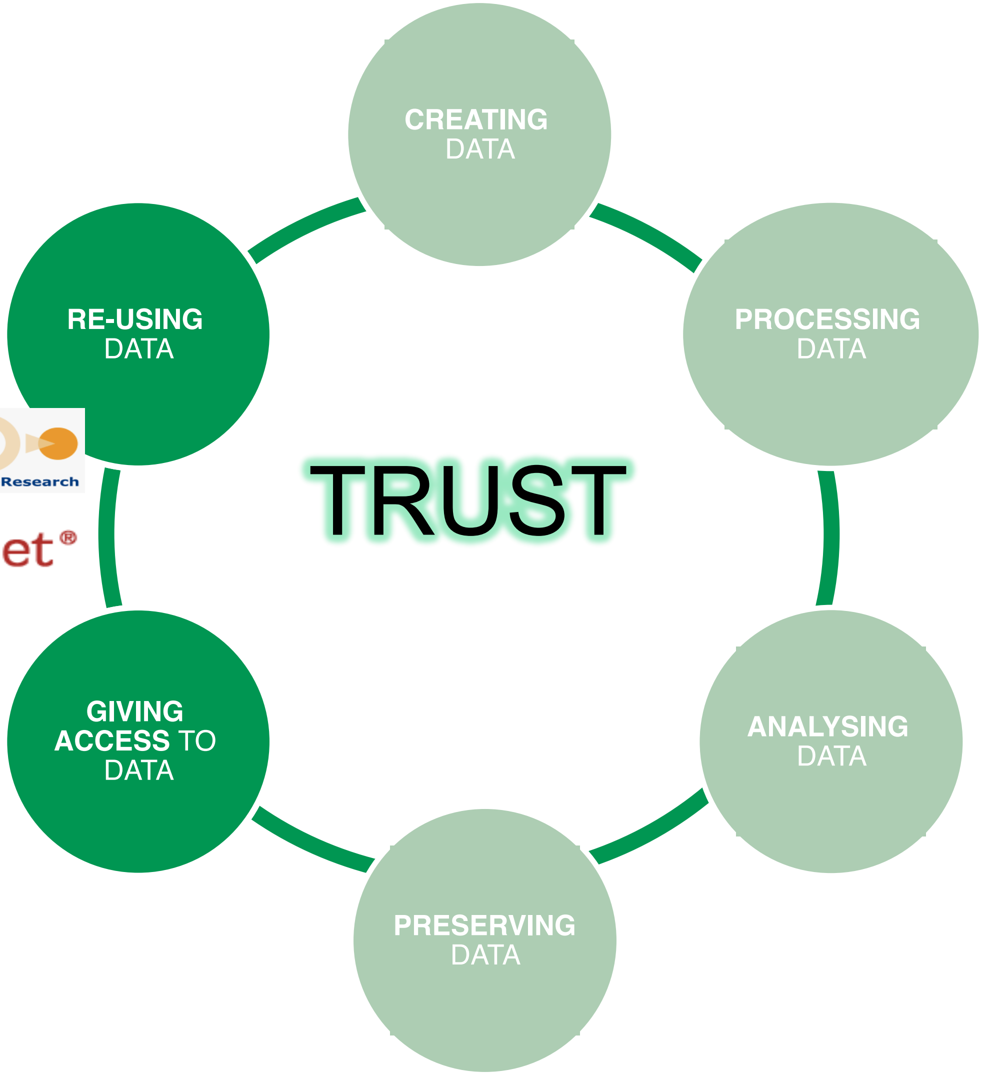
ePIC

Persistent Identifiers for eResearch

Handle.Net®



The data Life cycle



ePIC Persistent Identifiers for eResearch

Handle.Net®

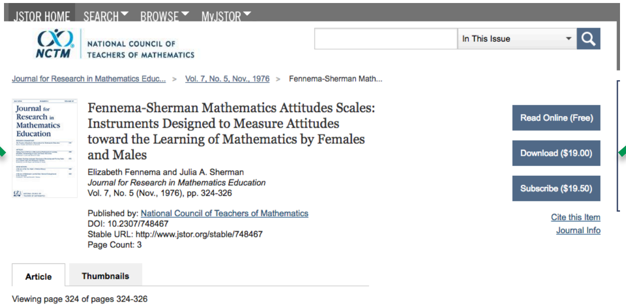
SURF SARA

PIDs

- PIDs (Persistent Identifiers) are
  - Pointers to resources like files, folders, webpages, real world objects
  - Globally unique
  - Resolvable via http

- Example resolvers: https://dx.doi.org/ and http://hdl.handle.net/

10.2307/748467



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

Journal for Research in Mathematics Education

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

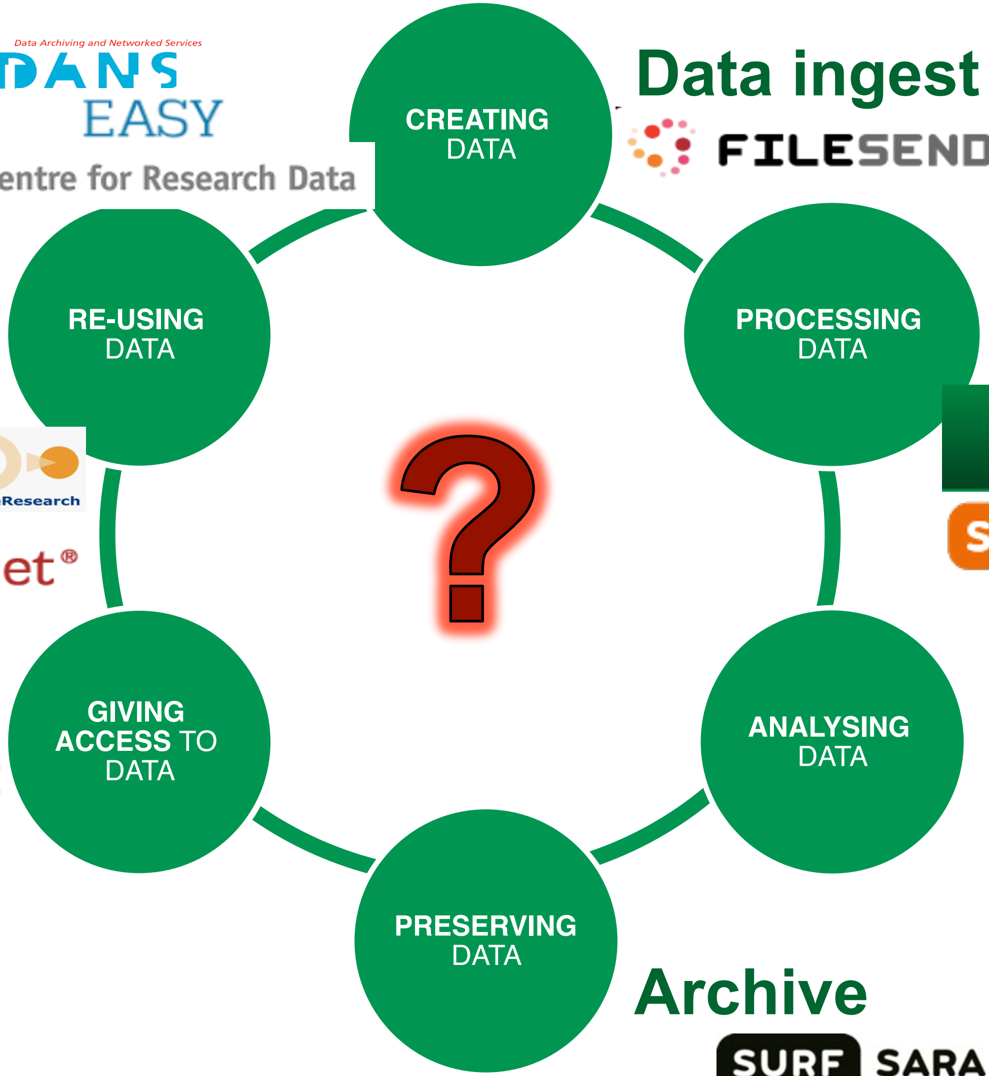
PIDs – Handle, EPIC and DOIs

- Handle
  - Technology to create, store and update PIDs
  - Infrastructure and technology to resolve PIDs
  - Example: https://dx.doi.org/ and http://hdl.handle.net/

- EPIC (European Persistent Identifier Consortium)
  - Maintaining reliable joint PID service for storing
  - Mirroring service, distributed resolving
  - Employing Handle technology

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

The data Life cycle



Data ingest service

FILESENDER

Data Archiving and Networked Services
DANS
EASY

4TU.Centre for Research Data

Data Archiving and Networked Services
DANS
NARCIS

B2FIND

ePIC
Persistent Identifiers for eResearch

Handle.Net®

BeeHub

SURF DRIVE

BeeHub

SURF DRIVE

Archive

SURF SARA