Making sensitive data FAIR

DANS Open Day
Open data, open science

Balancing access and protection

Joep Bos-Coenraad (PEP)
Ricarda Braukmann (DANS)
Cees Hof (DANS)

June 11 2026

DANS Radboud Universiteit

Programme for this session

13:00 - 13:10 Welcome & Introduction

13:10 - 13:30 PEP - Responsible Data Sharing

Repository

13:30 - 13:45 Looking forward & Discussion

Sensitive data - what do we mean? [1]

"Information that **may be regulated by law** due to possible **risk** for plants, animals, individuals and/or communities and for public and private organisations [...] All sensitive data are data that could **potentially cause harm** through their disclosure."

[RDA Sensitive Data Interest Group Charter]

Sensitive data - what do we mean? [2]

"Research data that **cannot be openly shared** for **some reason**, for example because it contains personal or confidential information or because there are copyright or other legal restrictions, and therefore **requires controlled access** or additional **safeguards**..."

[CESSDA Sensitive Data Group]

We want sensitive data to be FAIR

Sensitive data are extremely interesting for research!

- Addressing youth mental health issues
- Understanding long-term unemployment
- Studying migration patterns and integration processes

Various data sources exist in the Dutch landscape

- CBS microdata [various data linked to individuals, businesses and addresses]
- FIRMABACKBONE [information on Dutch companies]

How to make them FAIR?

Making sensitive data FAIR

Recognized as a major challenge in the SSH domain

Key bottleneck areas

Summary of challenges

- Increasing the amount of findable, accessible, interoperable and reusable (FAIR) research data and software
- Raising awareness amongst researchers about FAIR data and software practices
- Enhancing the awareness of available digital data and SSH-oriented tools
- Addressing pressing issues related to the collection and usage of data, like privacy and copyright legislation and the high costs of collecting and producing digital data
- Building an open, inclusive and equitable network

5. Sensitive data

A peculiar aspect of dealing with data is privacy legislation and how it is applied at many locations. Many researchers are of the opinion that the interests of the research unnecessarily suffer heavy damage. The social sciences, in particular, struggle with this, but it also increasingly applies to the humanities, where social media data are an increasingly important source for research. The necessary legal knowledge and support are often lacking and, insofar as it is present, **university legal advisers often prefer to take no risks**, as a result of which entire research projects must sometimes be modified or even called off. **Dealing with personal data often requires special digital environments, to which not all researchers can easily obtain access, insofar as these are present in the first place.** Another problem is found in the copyright legislation, which can throw up expensive obstacles in the case of media studies, art history or literature studies, for example. Our consultations revealed that various universities are working on digital tools that state whether an action is permissible. The TDCC could probably play a role in helping to realise collaboration and alignment in this area.

https://tdcc.nl/https://tdcc.nl/wp-content/uploads/2023/02/Roadmap-TDCC-SSH.pdf

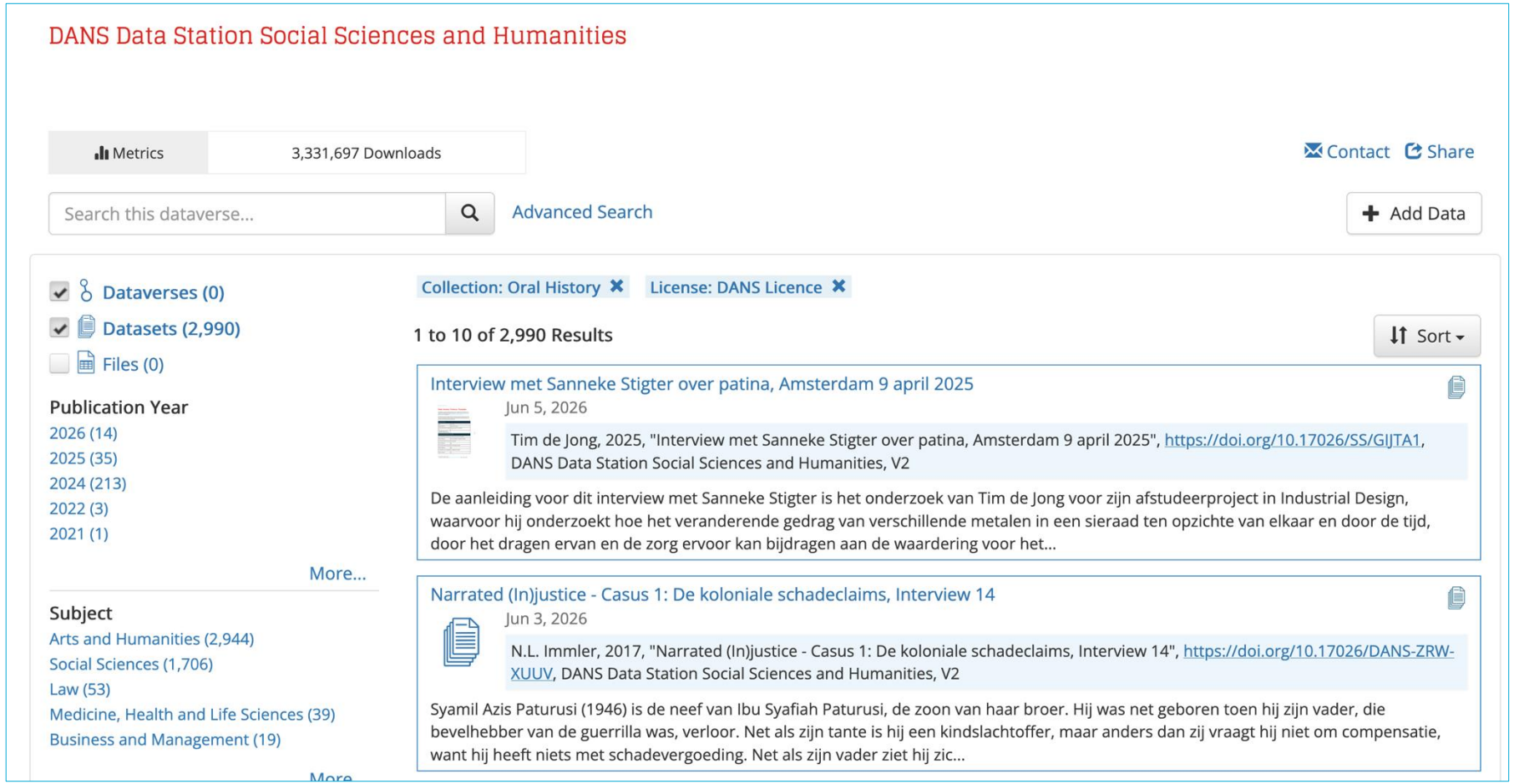
The current landscape for FAIR sensitive data

- No overarching standard process available:
  - Procedures and available infrastructure vary from institute to institute

- Elements for different aspects of the life cycle are available
  - **Storage during research**: Local storage solutions (e.g. PEP)
  - **Sharing for data analysis:**
    - Local storage solutions (e.g. PEP)
    - Secure analysis environments (e.g. SANE, CBS microdata environment)
  - **Long term archiving and publishing**: …

DANS Data Stations

- Allows the deposit of **restricted access** data



DANS Data Stations

- Allows the deposit of **restricted access** data

- Restricted access ~= sensitive data
  - Online repository system designed to make data (easily) reusable
  - Once access is granted data can be downloaded on any device
  - DANS staff has access to the data to verify and curate
  - No standard encryption of data

→ Not suitable for sensitive data of which disclosure could cause harm

PEP

Responsible Data Sharing Repository

Three methods to work with sensitive/personal data

- Bringing the question/algorithm to the source
  - e.g. Personal Health Train (PHT)
  - Decentralised data analysis
- Multi Party Computation
  - Analyse sensitive data using homomorphic encryption
  - All parties involved in decrypting results
  - Limited set of operations possible
- Data Sharing
  - Researchers are granted access to personal data
  - Preferably via a secure environment (e.g. MyDre or SANE)

No ‘one size fits all’ solution

- Keeping (unencrypted) data at the source is great…
  - … but not always applicable
- Whatever method is chosen:
  - Consider auditability, reproducability, etc.
- Secure archiving/managing/sharing of data?
  - The Open Source PEP Repository!

What is the PEP Repository?

- Responsible Data Sharing
  - **P**olymorphic **E**ncryption and **P**seudonymisation
  - Manage access and pseudonym translations for different users
  - Combine pseudonimised data from multiple sources
    - without sharing data or identifiers among sources before sharing
    - linking possible on sensitive identifiers without exposing them
  - Built on the principles of Privacy by Design (Ann Cavoukian)
  - Prevents breaking contextual integrity (Helen Nissenbaum)

Contextual Integrity

- Privacy has significance in context

- We live in a natural way in different contexts
- We keep information in its context

- Privacy is not an absolute concept,
  is not the same as secrecy,
  not exclusively related to data protection,
  and is not just a personal thing

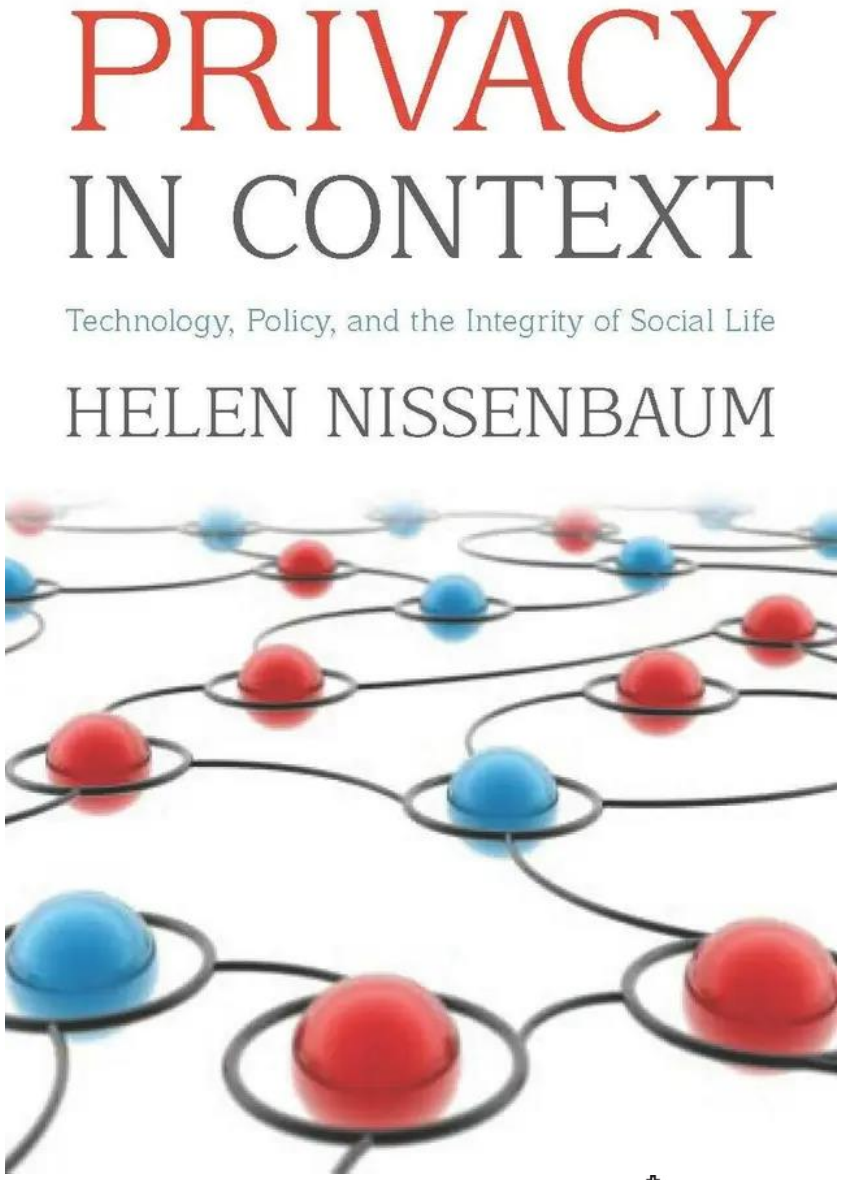


Exhibit A


BREAKING
Informateur Wijers legt functie neer om uitspraken in privéappje

Exhibit B

Exhibit C: Hack Clinical Diagnostics

- July 2025
- Cervical Cancer Research
- 485 000 voluntary study participants
  - Names, Addresses,
  - Social Security Numbers
  - Phone numbers
  - Medical Details

Approach on data protection is evolving

- Currently, data protection mostly consists of:
  - Protect against externals

- What we actually need (also according to the GDPR)
  - Proportionality
  - Minimisation
  - Subsidiarity
  - Pseudonymisation

Protect against who?

Protection against... (1/4)

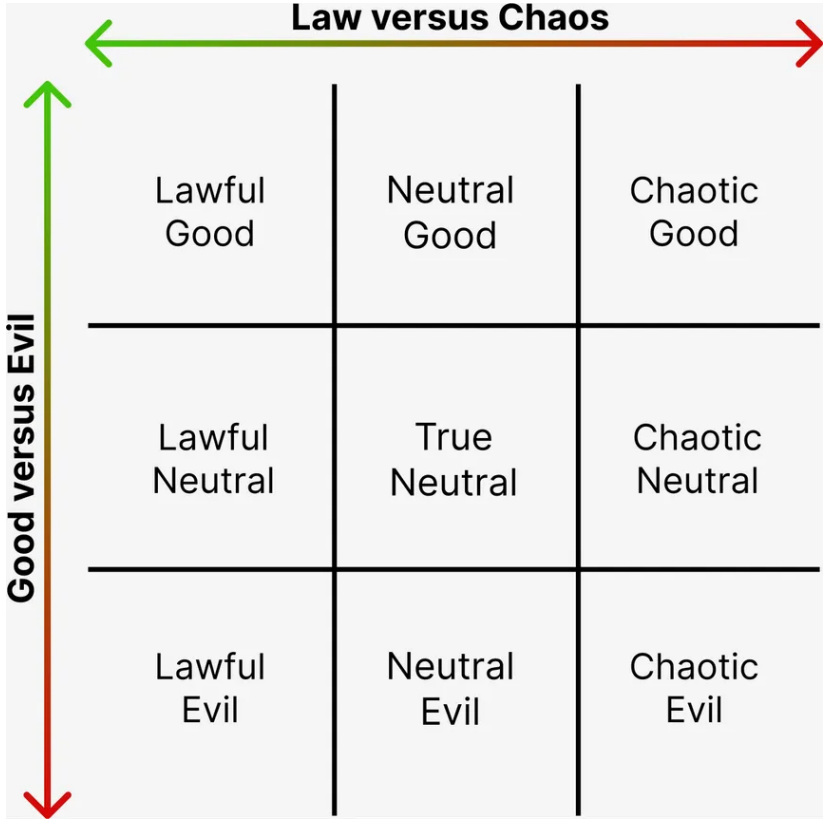
The unauthorised malicious actors (lawful-chaotic/evil)

- Shouldn't access any personal data.

Examples:

- Geopolitical actors
  - disruption/stategic theft
- Commercial or political exploitation
  - unsolicited marketing & targeting / influencing politics / etc.
- Scammers, thiefs (@see fig.)

Protection against... (2/4)



The researcher(s) (neutral-chaotic/good)

- Requires data
- Can not always work with anonymous data
  - (requires pseudonymous data)
- Must work with minimised datasets
- Must respect proportionality and subsidiarity
- Should break down/proxy analysis (Partial Data Analysis)
  - Lower source data entropy with siloed preprocessing
  - For persons and systems
- Reproducibility, auditing

Protection against… (3/4)

The sysadmin(s) (lawful/neutral, sometimes BOFH)

- System depends on sysadmins (manage ops, maintain uptime, backups)
- Shouldn’t access unencryted personal data

Protection against… (4/4)

The dataset access manager(s) (lawful/good)

- Managing access for persons or systems to a dataset
- Requires access to metadata
- May require access to low-entropy unencrypted dataset contents
- Shouldn’t access additional unencryted personal data

Privacy Enhancing Techniques in PEP (1/2)

- End-to-end-encryption
  - Only Data Sources and End Users get access to unencrypted data
  - Data is encrypted **in transit** *and* **at rest**
  - Servers nor sysadmins are able to decrypt
- Access Management
  - Fine grained read/write access management, also for metadata
  - Data minimisation (GDPR requirement)
  - Dataset managers don't require access to unencrypted data either

Example: Data Management

Example:
A Researcher requires read access to some survey and application data for a defined cohort (group of subjects)

After analysis, derived data is written to the ‘Derived Analysis’ column using the researcher’s local pseudonyms

Columns →	Contact Info		Survey 1		Survey 2		Application Data		Genome Data		Live Sensor Data		Derived Analysis		MRI Scan	
Data Subjects ↓	Read	Write	Read	Write	Read	Write	Read	Write	Read	Write	Read	Write	Read	Write	Read	Write
Subject 1			✅ read				✅ read							✅ write		
Subject 2																
Subject 3																
Subject 4			✅ read				✅ read							✅ write		
Subject 5																
Subject 6																
Subject 7																
Subject 8			✅ read				✅ read							✅ write		
Subject 9			✅ read				✅ read							✅ write		
Subject 10																
…																
Subject N-1			✅ read				✅ read							✅ write		
Subject N																

Privacy Enhancing Techniques in PEP (2/2)

- Four eyes principle 👀 👀
  - New data access is granted via two independent roles
- Advanced Key Management
  - Complementary encryption keys at different ISP's
- (Micro)Pseudonymisation
  - Only PEP Repository can link (encrypted) pseudonyms to records
- Data Versioning
  - Users no longer keep dataset backups

Why this level of security?

- The sensitivity of the same data has increased over time.
  - Nowadays, data snippets are easily collected, combined and exploited

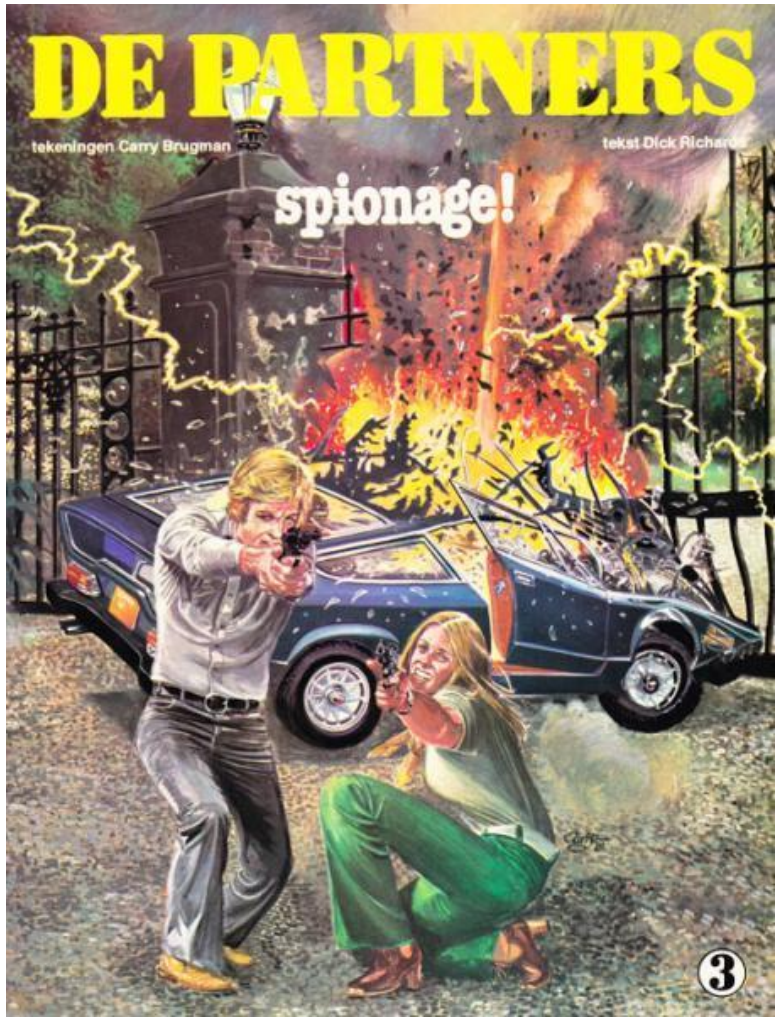
- Commercial and criminal hunt on personal data peaks.

PEP Partners

- Personalised Parkinson Project (PPP)
- Healthy Brain Study
- NOLAI (National Education Lab AI)
- UvA

Exploring partnerships with:

- SURF
- Erasmus University
- ProcoliX
- DANS



PEP Team:

Founding members / advisory team

Bart Jacobs is professor of computer security at Radboud University and the scientific leader of the PEP project.

personal website

Jean Popma is project manager of the PEP project.

LinkedIn

Operational team

Hans Harmannij is a Scientific Programmer. In the PEP project he works on authentication, access control, and interfacing with software and procedures used by research groups.

Kai van Lopik is a scientific programmer with a background in commercial software development. Focusing on usability and maintainability, he strives to make PEP as simple as possible (but no simpler).

LinkedIn

Joep Bos-Coenraad is project leader and developer at PEP. Aside from software development Joep has a background in physical sciences, academic research, politics and entrepreneurship.

Personal page

Steven Wallis de Vries is a software developer at PEP since Oct. 2023, with a background in computer science and cyber security.

Sander Meis has a background in material science physics and software development. Sander is a skilled scientific software developer dedicated to advancing research and innovation through the design, implementation, and optimization of cutting-edge software solutions.

Linus van Elswijk is a software developer at PEP, with a background in computer science and game & media technology. His focus is code quality and testing, with the goal of keeping the codebase flexible and open to change.

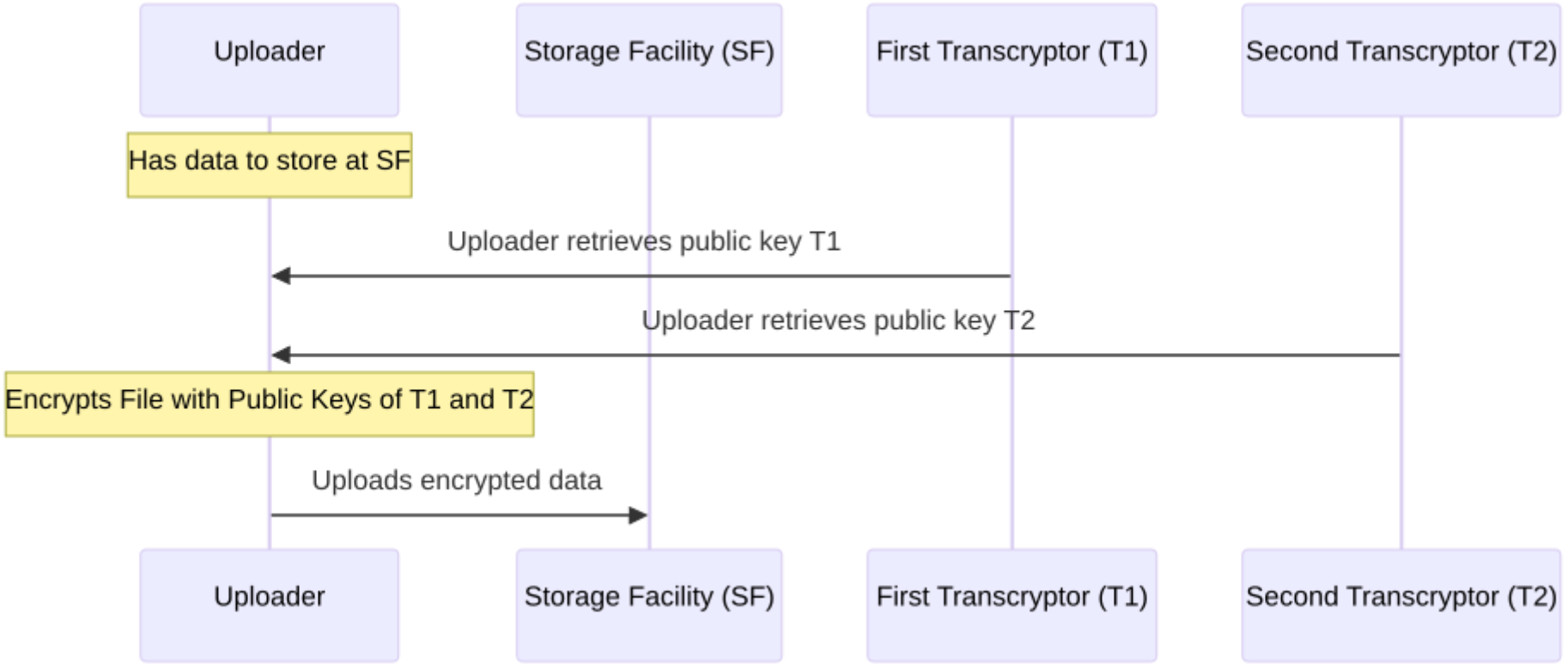
PEP Website:

- https://pep.cs.ru.nl/

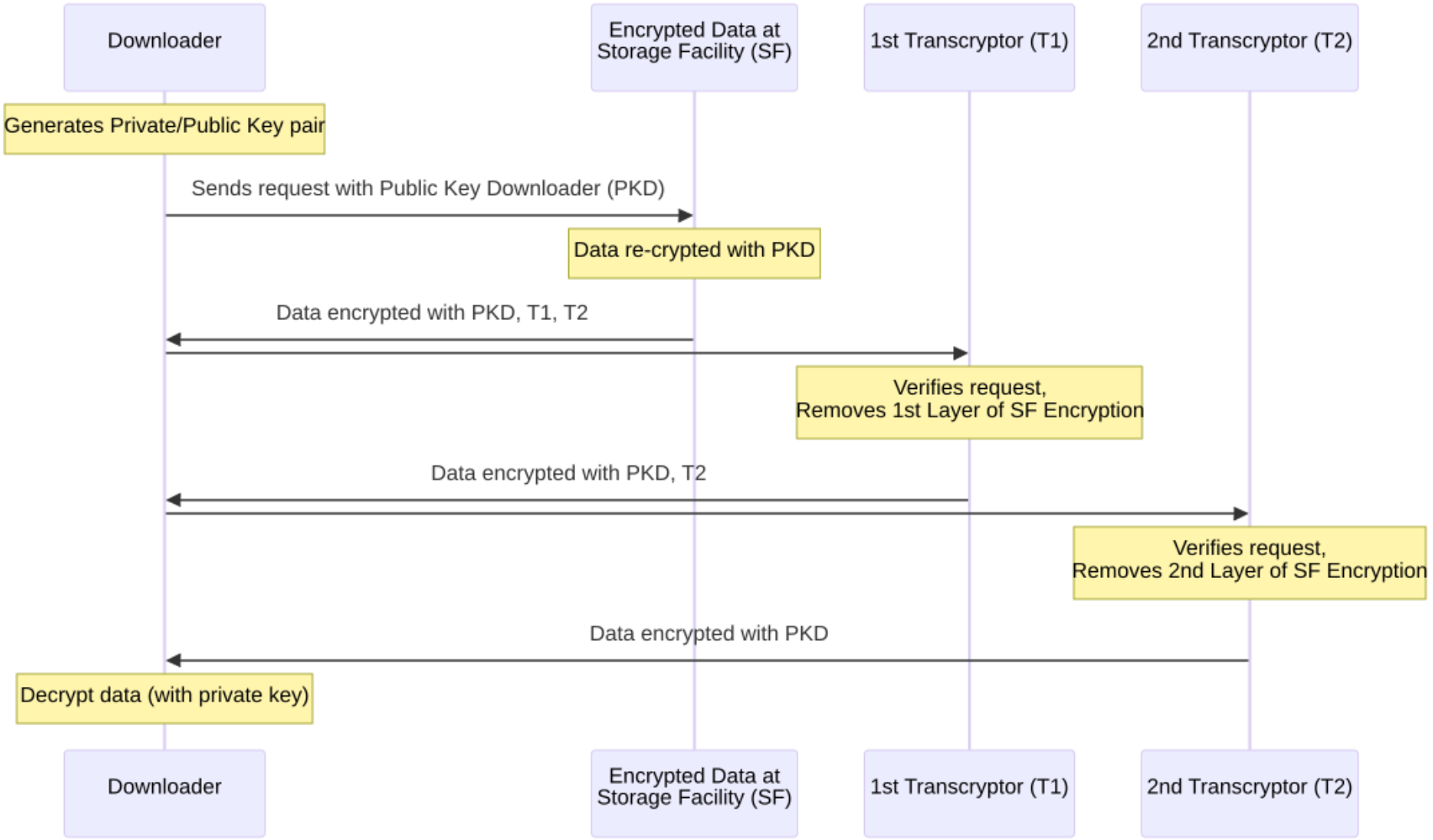
PEP Repo source on Github:

- https://github.com/pep-repository

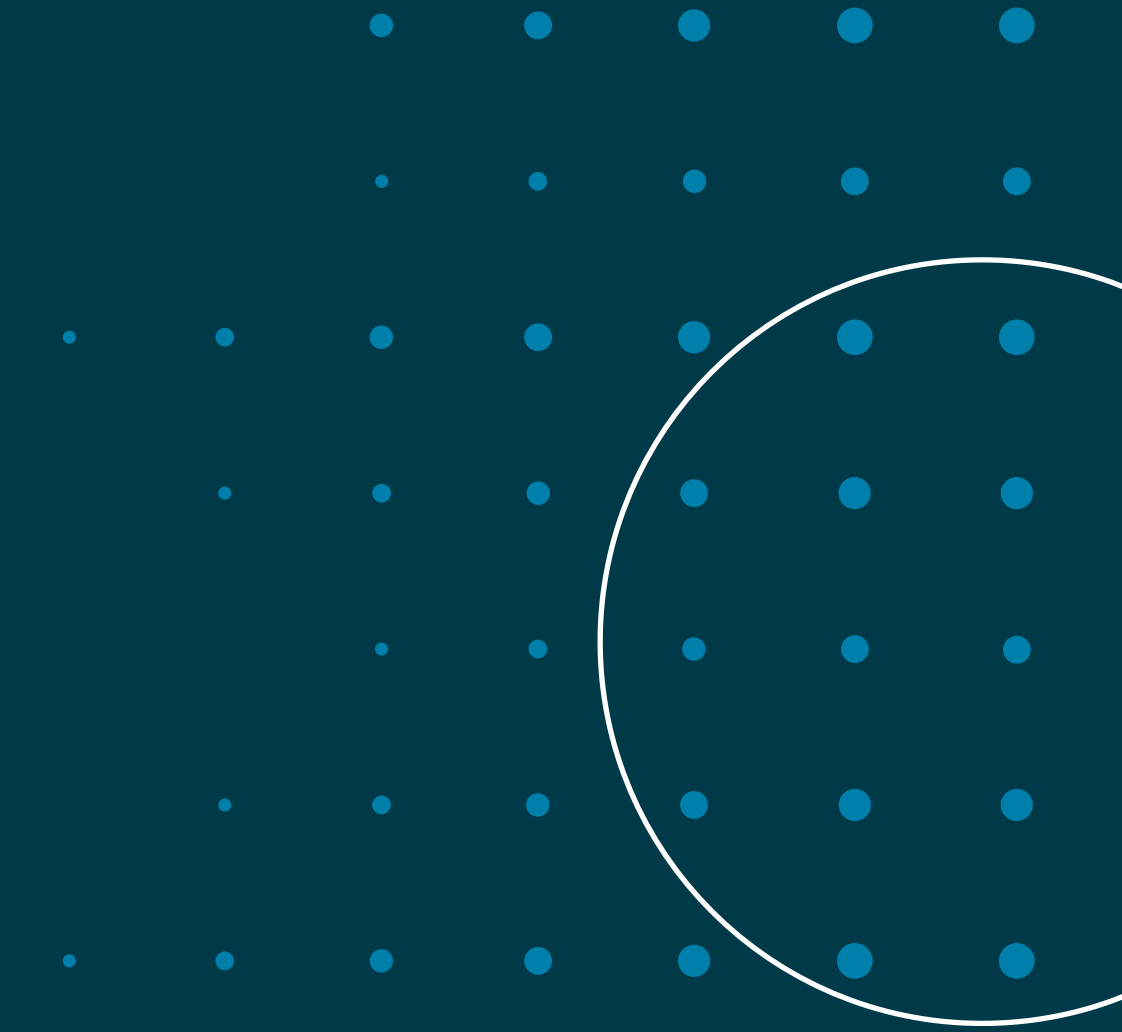
Example: uploading data to PEP Repo



Example: downloading from PEP Repo



Looking Forward

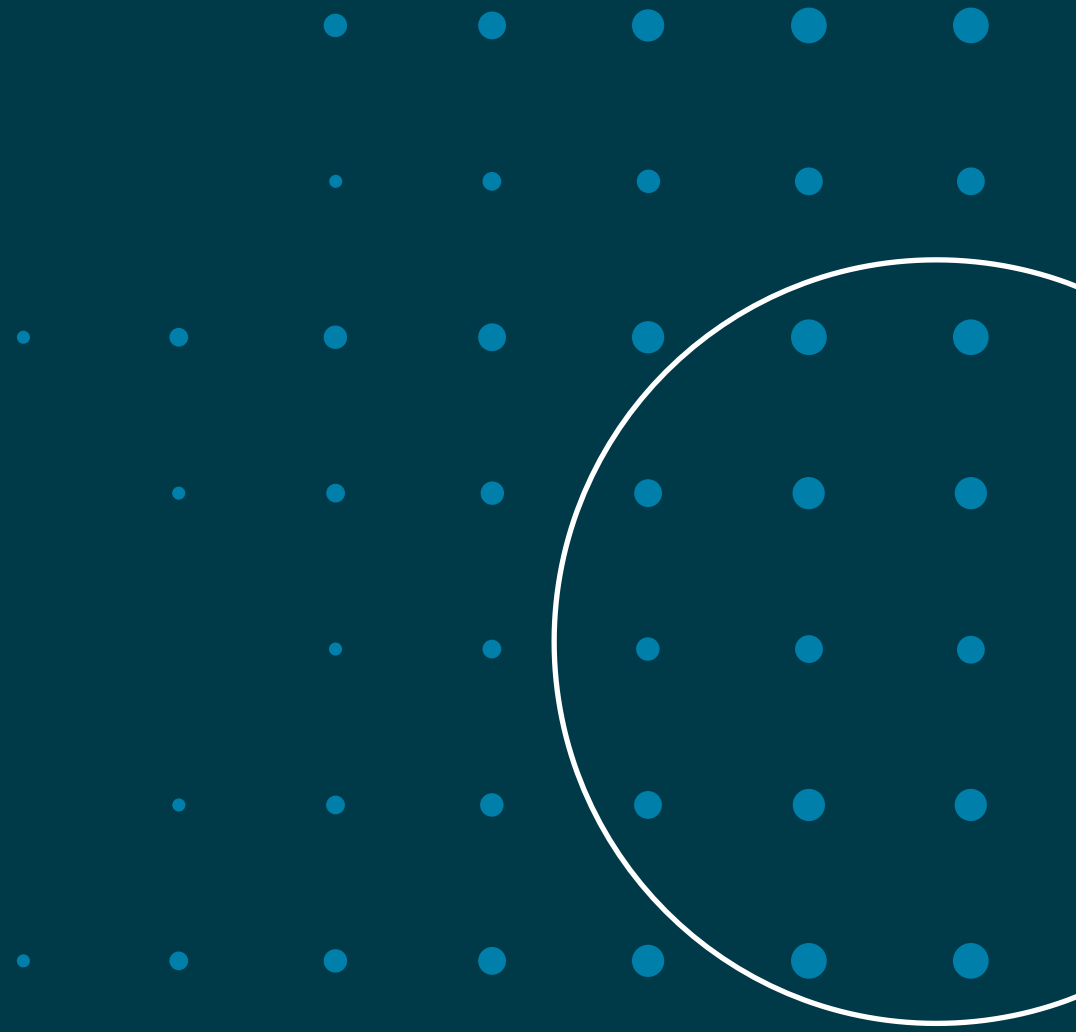


Towards making sensitive data FAIR

- Connecting the tools that are already available
  - Ensuring the different (local) solutions can integrate

- Build upon the **strengths** of the different systems
  - PEP : Local storage and secure access to data
  - DANS: Making metadata findable
    - Providing a sustainable connection to where data are stored
  - DANS: Open access to documentation, non-sensitive data
    - De-identified derivatives
    - Synthetic data

Discussion



Thank you for your attention

DANS Open Day
Open data, open science

DANS

Anna van Saksenlaan 51 | 2593 HW The Hague | The Netherlands | +31 (0)88 003 46 66
National centre of expertise and repository for research data | An institute of the KNAW and NWO

DataLink LinkedIn BlueSky Mastodon www.dans.knaw.nl