

GUIDEBOOK

# Hard-to-share data

in the Social Sciences and Humanities

Part of the project

'Beyond Personal Data: RDNL training on hard-to-share data for SSH early-career researchers'





# Table of Contents

<b>Attribution</b>	<b>4</b>
Funding	4
Authorship	5
<b>Introduction</b>	<b>6</b>
About the project	6
About this guide	6
Hard-to-share data	7
Personal data	7
Non-personal data	8
<b>Who is involved in the management and sharing of hard-to-share data in SSH</b>	<b>13</b>
Insights from data stewards, ethics facilitators, and privacy officers	13
<b>Case studies of 'hard-to-share' data in SSH</b>	<b>17</b>
Case Study 1: Navigating Hard-to-Share Data in Corporate Crime Research	17
Case Study 2: Sensitive Qualitative Interviews	19
Case Study 3: Sharing location data: Sensitive data in archaeology	21
<b>A solution for sharing hard-to-share data in the SSH: Secure ANalysis Environment (SANE) for inspecting and analysing data without downloads</b>	<b>23</b>
An example of SANE in action: the FIRMBACKBONE project	24
<b>Final thoughts</b>	<b>27</b>
<b>Training materials from the three-part workshop</b>	<b>29</b>
<b>References</b>	<b>31</b>

# Attribution

## Funding

This publication is part of a series organised by the project 'Beyond personal data: a new initiative to support early-career researchers with hard-to-share data' financed by the Dutch Research Council (NWO) via the Thematic Digital Competence Centre Social Sciences & Humanities (TDCC-SSH)

## Licence





## Authorship

This guide has been written by:

**Deborah Thorpe,**

DANS

<https://orcid.org/0000-0002-2307-8770>

**Michelle van den Berk,**

DANS

<https://orcid.org/0000-0002-1218-8448>

**Pascal Flohr,**

Leiden University,

Leiden University Libraries

<https://orcid.org/0000-0003-3203-913X>

**Marjan Grootveld,**

DANS

<https://orcid.org/0000-0002-2789-322X>

**Ahmad Hesam,**

SURF

<https://orcid.org/0000-0001-7331-1000>

**Lucas van der Meer,**

ODISSEI

<https://orcid.org/0000-0003-4415-678X>

**Maithili Kalamkar,**

SURF

<https://orcid.org/0000-0003-4378-1612>

The authors would like to thank the following contributors:

- Jing-Yi Magraw, Emilie Kraaikamp, and Myrte Vos for their advice, which has greatly enriched the section: 'Who is involved in the management and sharing of hard-to-share data in SSH'.
- Rebecca Campbell and Frederike Oberheim for the informative and highly interesting case studies that they presented during the original training workshop, which provided source material for this guide.
- Shoko Dauwels for attending the training workshop as an observer and providing thorough and useful feedback on that workshop, which we have tried to address in this guide.

Importantly, we would like to thank the participants of the workshops that we organised as part of the Beyond Personal Data project.

Cover image: Freepik. [www.freepik.com](http://www.freepik.com)

January 2026

DOI: <https://doi.org/10.5281/zenodo.17588795>

# Introduction

## About the project

Sharing research data as openly as possible is key to responsible and reusable research: it increases the transparency of research and increases the chances that data will contribute to future studies. One of the main bottlenecks to sharing research data in the social sciences and humanities (SSH) is that many datasets contain personal data, which legally and ethically cannot be shared openly. Consequently, attention has been mostly focused on the challenges posed by these sorts of datasets. While protecting personal data from unwanted disclosure is a crucial issue, there are many other reasons why data can be hard to share that have not yet been sufficiently addressed.

In response to this, a short project entitled 'Beyond personal data: RDNL training on hard-to-share data for SSH early-career researchers' was initiated by project partners DANS, SURF, Erasmus University Rotterdam (International Institute of Social Studies as well as ODISSEI), Leiden University, and the Promovendi Network Nederland (PNN), and funded through the Dutch Research Council (NWO) via the Thematic Digital Competence Centre Social Sciences & Humanities (TDCC-SSH). The project ran for 12 months during 2025, and this guide is one of the outcomes.

As the scope of non-personal hard-to-share data is wide, we identified three specific

bottlenecks to address during the project, chosen based on current gaps in training and guidance, as well as what could be feasibly addressed in this project:

- Sharing non-personal sensitive data that cannot be openly accessible, including the solutions offered by the SANE (Secure ANalysis Environment)<sup>1</sup>
- Sharing field notes
- The ethics of sharing fieldwork data

For each of these topics, a half-day in-person training workshop was organised, of which the training materials have been published (see p.25 below). In order to extend this work of helping researchers to navigate this complex topic, we also proposed to write a guide on each topic; you are currently reading the first guide.

## About this guide

In this guide, we focus on sharing non-personal sensitive data. We refer to the word 'sharing' primarily in reference to 'giving access to others': ranging from sharing data with colleagues to publishing in a data repository. However, sharing can also refer to activities that involve *taking access*, for instance analysing data that have been generated by others; for that aspect we include, towards the end of this guide, a section on the use of the SANE secure environment for accessing data safely.

<sup>1</sup> <https://www.surf.nl/en/themes/research-infrastructure/sane-secure-environment-for-analysing-sensitive-data>

Sensitive data can often not be made openly accessible or even just shared with, or sent to, other researchers. There are many possible reasons for this, for example because there are commercial or security constraints. In this guide we focus on exploring the types of non-personal sensitive data, as well as providing some case studies with possible solutions. One of these solutions, SANE, is introduced in detail.

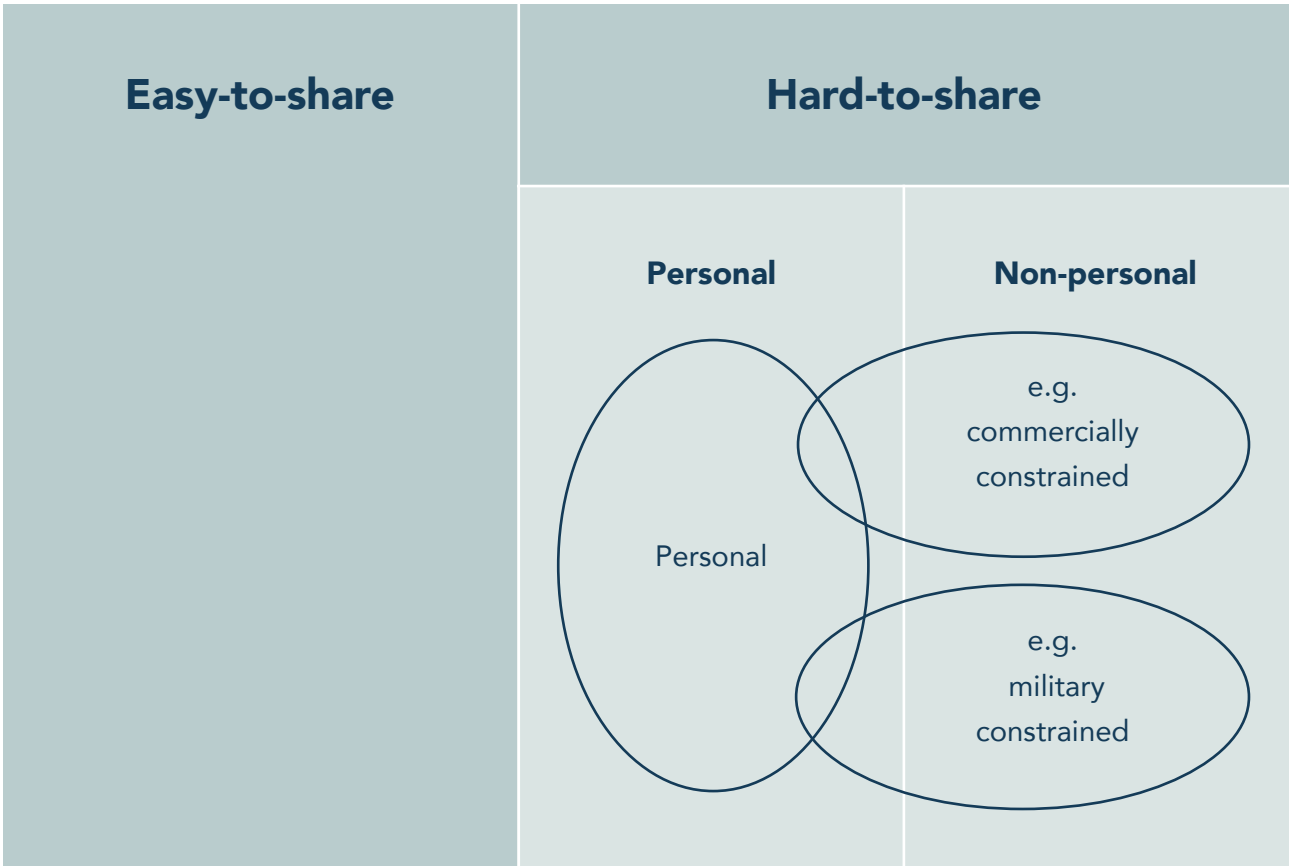
The guide, to be used as a standalone source of information, is based on the first ‘Beyond Personal Data’ workshop, of which the training materials are also openly available (Thorpe *et al.*, 2025).

### Hard-to-share data

There are several types of research data that are considered to be hard to share in the SSH. In this section, we provide more information about these types of data and why they are challenging to share, beyond the original context in which they were collected.

### Personal data

First, we must take a sidestep back to personal data. In the SSH, researchers frequently work with data that can contain some personal data, such as data collected through surveys; interviews; or archival materials (e.g. diaries, letters, and photos which may contain information about still-living individuals). Due to privacy concerns, these cannot be shared openly without data processing and sharing



precautions. Because sharing personal data responsibly is, for many researchers in these domains, the most frequently-encountered data-sharing challenge, a lot of attention has been paid to this in existing training and educational resources.<sup>2</sup> A 2024 webinar focused on ‘non-personal’ sensitive data, with some of the same authors as this guide. Readers may find it useful to consult the slides and recording (see Grau *et al.*, 2024).

However, we observed in the process of offering training on ‘non-personal’ sensitive data that there is, nonetheless, a need for even more training and guidance on the topic. Therefore, we provide below some key information and resources to consult when sharing personal data.

Personal data are any information that can be used to identify a living person. It can be both obvious, directly identifiable, data such as someone’s name or address, but also indirectly identifiable data, in which case it is often a combination of factors, like ‘the Law Faculty data steward’ or ‘the only psychology student at Leiden University who is 67 years old’ that makes the data identify an individual. Especially in the social sciences, such as in the health sector, a lot of research is done with living human subjects, often through interviews or surveys, and these datasets are likely to contain personal data. Since the General Data Protection Regulation (GDPR) came into force in the EU in 2016, there are legal restrictions on sharing personal data that apply across

the EU. Of course, long before the GDPR, researchers in the EU as well as globally have been observing national rules and ethical guidelines to protect their research subjects.

While it is more challenging to share datasets that contain personal data – especially to make them openly available – there are various solutions for deidentifying the data. Anonymisation or pseudonymisation, when properly applied, can make it impossible or much more difficult to identify a participant; if this is somehow not possible, then a last resort can be to try to publish at least the metadata and, if possible, some spreadsheets that contain unidentifiable data at an aggregated level. In the case of research on non-sensitive topics, research participants may also be happy for their pseudonymised data to be shared by means of transforming personal data in such a way that it can no longer be traced to a person without additional information. In some cases, research participants may even prefer to remain identifiable, for instance with research involving creative practitioners who may be proud to be recognised.

When gaining informed consent to process personal data, information should be given to the participant specifying exactly how the data will be processed and, if relevant, shared. If permission for other researchers to work with the data has been obtained through informed consent, a secure analytical platform like SANE may be a good solution for granting access.

---

2 See, for example the 2022 webinar, ‘FAIR for Sensitive Data’, which focused on ‘introduced tools that can be used for data anonymization, privacy-preserving processing, generation of synthetic data, etc’ (see Jaber, 2022).

## Key resources on the use of personal data in SSH

- A practical guidebook for researchers and data stewards that provides an overview of how to make qualitative data reusable created by a team at DANS (see Verburg, Braukmann, and Mahabier, 2023).
- CESSDA's 'Data Management Expert Guide', with a useful section dedicated specifically to anonymising data (see CESSDA Training Team, 2017 - 2022)
- A blog post that explores the delicate balance to be achieved between the need and obligation to protect research participants and their personal data, and the importance of maintaining the utility and reusability of your data (Thorpe and Braukmann, 2024).
- A guide with a focus on pseudonymisation, specifically, as a subset of data deidentification (Kleef *et al.*, 2019)
- An article that focuses on two real worked examples of de-identification of data in the behavioural and social sciences, exploring 'how to be as open as possible and as closed as necessary with the goal of maximally facilitating science while minimizing the risk of participant identification' (van Ravenzwaaij *et al.*, 2025)

## Non-personal data

While many datasets in the SSH contain personal data, there are other reasons why it may be more difficult to share data in these domains. These reasons require different approaches and solutions to handling and sharing the data responsibly. There is a broad group of themes that fall under this category, some might overlap, and it is not possible to be absolutely complete here. However, in this section we attempt to give an overview of the most common and pressing types of non-personal hard-to-share data that are relevant to the SSH domain. In terms of types of data, these comprise both qualitative and quantitative data which can be observed, collected, or experimental, and include many data formats.

Common reasons for data being hard(er) to share (not including privacy/personal

data protection):

- Commercial constraints
- Security constraints, including due to knowledge security
- Ethical considerations
- Risk of harm to subjects of research
- Challenge in sharing non-structured analogue data
- Legal constraints

In addition, there may be practical issues like a lack of a technical solution to store, share, and find sensitive data. It is important to recognise that even if data cannot be published or otherwise shared openly, **it is still almost always possible, and important, to share FAIR metadata openly.** By doing this you are making your best effort at FAIR-ifying the dataset by making others aware of its existence and providing information about the research. The only exception to this may be highly security-constrained research data. Below we

provide more information about the types of data listed above:

**Commercial constraints:** For example, financial data that could be useful for research in economics or law, but are restricted because of commercial interests. Other examples might include information about how commercial businesses are operated (i.e. in management studies) or analyses of commercial products from the film or music industry (e.g. in film studies). Beyond the SSH, another example are pharmaceutical or chemical data, which may lead to patents (be it for the research group, i.e. university, or for a commercial funder).

*Ways to start addressing the challenges:*

- Clear, formal agreements are important here for (the duration of) data access, data transfer, and the availability, openness, and 'ownership' of and responsibility for resulting collated or aggregated data, as well as an authorship.
- Secure environments such as SANE could be used to share data for the purpose of analysis.

**Security constraints, including knowledge security:** For example military information which could be useful for research in governance or global affairs studies, but that needs to be protected from unwanted access and could pose a risk to national security. Criminological information which could be used by hackers or for other criminal purposes. Other data that poses risks of misappropriation of knowledge and/or terrorism.

*Ways to start addressing the challenges:*

- Comprehensive risk analysis before and clear

agreements with stakeholders on what can and what cannot be shared (see for example: Government of the Netherlands, n.d).

- If there are security constraints, secure environments could be used for data sharing.

**Ethical considerations:** this is partly also an issue of 'knowledge security', you may need to protect data which could be misused by countries or organisations involved in human rights abuse. This also includes consideration of the question who 'owns' the data and who should get a say in how it is used. For example, data derived from fieldwork in which local communities are involved or are present in the area. Furthermore, this also includes data about people that are no longer alive: while GDPR does not apply to deceased people, there may still be reasons why the data are sensitive, for example to protect their descendants. Ethical considerations around data sharing is a very broad area, and you can start by searching for resources that are specific to your own research domain.

*Ways to start addressing the challenges:*

- Familiarise yourself with the literature on ethics in your field.
- Familiarise yourself with the CARE Principles for Indigenous Data Governance (Carroll et al., 2020).
- Start from the 'Do no harm' principle (International Red Cross, 1965)
- Consider who you should involve in thinking about and deciding on sharing data.

**Risk of harm to the (non-human / non-living-human) research subject:** While GDPR and ethical considerations take into account possible

harm to (living) persons, we should also consider potential harm caused by data sharing to animals, plants, areas, objects, and more. Examples are the sharing of location data which may cause harm to cultural heritage (archaeology, history) or, mostly beyond the SSH domain, the natural environment (geology, ecology).

*Ways to start addressing the challenges:*

- Assess the risk and impact of the harm and the sensitivity of the object to be harmed by the data release (Chapman, 2020). Depending on the outcomes, decide if, and at what resolution, the data should be openly shared.
- Depending on the field and type of data, there will be various options to obscure parts of the data or make them less precise, for example the case study on location data below.

**Challenge in sharing non-structured, analogue data:** Examples are written notes and sketches, like field notes (e.g. in archaeology, cultural anthropology/ethnography, sociology).

*Ways to start addressing the challenges:*

Sharing these types of data is essential for research integrity and the verification of results in the aforementioned fields. However, they are diverse in their type and often not shared for various reasons. These reasons are complex, as is the guidance and best practices that may help to address them. With this in mind, our project team has written a dedicated guide entitled 'Sharing Field Notes' (Flohr and Roodhof, forthcoming).

**Legal constraints:** Of which GDPR is one, but there are also other legal considerations

to take into account when collecting, using and potentially sharing data, ranging from copyright to local laws on sharing heritage data. For example, datasets that comprise data that are owned by others, like private archival materials or copyrighted texts. Legal constraints may also coincide with commercial constraints.

*Ways to start addressing the challenges:*

- Seek advice from an expert (for example a data steward and/or a privacy officer) and together examine potential legal aspects around sharing the data, for example protection of personal data, data ownership, contractual obligations, and licensing (see the section below, 'Who is involved in the management and sharing of hard-to-share data in SSH').

Each of these reasons, in itself, is broad, and existing and new solutions require in-depth discussion as well as an awareness of the specific context in the (sub)field – the above ways to address challenges are very generic and a starting point only. There are many, and diverse, ethical considerations around sharing field work data and field notes, and for these we refer to the two other guides created during this project (Flohr and Roodhof, forthcoming; Lushaj et al., forthcoming). In the current guide we focus on (some) possible solutions for sharing data with commercial or security constraints, with case studies from economics and law, respectively, and on sharing location data, with a synthesising case study from archaeology.





# Who is involved in the management and sharing of hard-to-share data in SSH

All researchers work with data, but not all researchers are, or should be, equipped to tackle the challenges involved in data management and sharing alone.

We acknowledge that not all researchers have access to the same resources in terms of advice, training and hands-on support. However, there are many types of skilled and knowledgeable professionals who are involved in managing and sharing hard-to-share data in the social sciences and humanities, including but not limited to:

- Other researchers who can provide domain-specific information and advice
- Research data management specialists or data stewards who might provide advice, training, and hands-on help with data management planning
- Privacy officers who provide insights into the legal issues around data collection, processing and sharing
- Ethics experts who are involved in the ethical approval process and can advise on the ethical issues around data collection, processing and sharing
- IT professionals at your institution who can advise and provide access to secure infrastructure

- Repository staff who can help and advise with data and metadata preparation and the process of sharing the data responsibly in a repository
- Other experts who provide training and education in research data management and FAIR data sharing.

In this section, we provide you with insights from a selection of these individuals.

We asked them for tips on how to deal with 'hard-to-share data' and we hope that you find them inspiring!

## Insights from data stewards, ethics facilitators, and privacy officers

### Jing-Yi Magraw,

Research Ethics Facilitator,  
Erasmus University Rotterdam

"In my role, I advise applicants more specifically on how to prepare for the ethics review process and their ethics applications. One of the most important aspects to consider is how to explain your methodology and your data to an institutional review board that may

not be familiar with your specific research project and field. For example, never assume that the reviewers are familiar with how you may choose to define a certain term or methodology that is integral to your research - in an ethics application, reflecting upon your methodology can include reflecting on how you yourself approach this particular method and its application to your project.

Another ethical consideration for non-personal sensitive data is for researchers to ask themselves: does my data/research have the potential to be misused? Unlike personal data, it may not be the case that non-personal sensitive data will directly impact a specific participant, but working with this data could have a broader, negative, impact on a community or the environment. How will you navigate and mitigate these potential risks of misuse?"

**Emilie Kraaikamp,**

Legal Support Officer and Privacy Coordinator,  
DANS

"'Hard to share data' in social sciences or humanities – or in any field for that matter – will always require special attention. The reasons why it is challenging to share certain types of data may differ considerably, however there are some things that you can take into account that will apply to all cases. For any type of 'hard-to-share data' a good preparation phase is key. My advice is to take the following three things into account.

If it is possible to select only those data that you need for your goal, make sure to select

only these. For instance: if you only need general information on an archaeological site and not its sensitive coordinates, leave the latter out. This will help you keeping the data that need protection to a minimum.

Think in stages when it comes to working with the data. Do you need the data in all phases of your project? Is it essential to share all data in the end? You may be able to replace certain data with more general information for a specific audience. If this works for your data you could create a public version and a sensitive version that is shared on specific terms.

The above always goes hand in hand with making the proper arrangement with the persons or organisations involved. Make sure that you have permission to achieve your goal. How to share the data shouldn't be an afterthought in this, but an integral part of this preparation. What part of the data can be shared, through which medium and under which terms? This can be lengthy process, where you may need to consult with multiple parties, but you will benefit from this the moment you are ready to share your 'hard-to-share' data."

**Myrte Vos,**

Data Steward,  
Leiden University

"Lots of what we might consider 'hard-to-share' research data in the Humanities falls under 'personal data' (though personal data does not always make research data hard to share!). Beyond that category, we can think



of data that is hard to share because it is very personal to the researcher themselves (think, fieldnotes or artistic practice), or because it falls under intellectual property rights, or because it should be governed by the community from which it was sourced (such as indigenous nations).

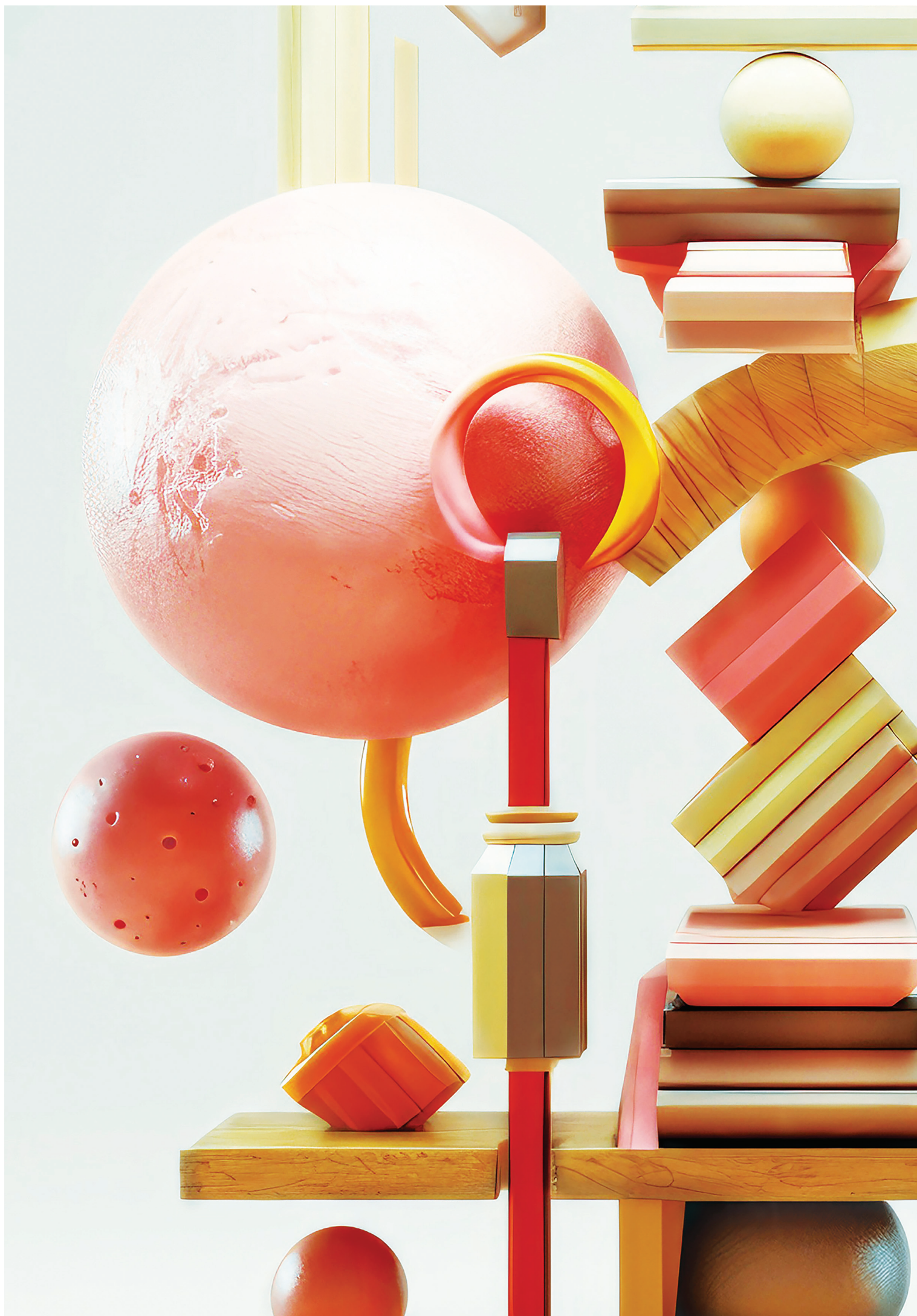
As data steward, I try to take the philosopher Sabina Leonelli's attitude to Open Science: we share data in order to form judicious connections.<sup>3</sup> Who does the researcher want to connect with? That informs the approach to sharing. The more research data seems to 'resist' straightforward sharing and publication, the more it invites us to be intentional about building those connections. That resistance is more than a nuisance: it really forces the researcher to think about what her data is and what it could effect, which is a deeply rewarding exercise.

The decision-making power over what, exactly, can be shared (and how, and when, and with whom) often lies with someone other than the researcher: the tribal council, the curators of the archives she draws from, the photographer she collaborates with, even her institution's own policies. But the responsibility to get it right, and make sure that promises to share are kept, lies with her. That is no light task, and simply not sharing can feel like the safer and easier option.

As data steward, I try to empower researchers to use the agency they do have – to lean on their expertise and experience,

the relationships they already have within communities and with research partners, rather than become paralysed with fear that they will violate some policy or law. I can help clarify which boundaries are hard, and which are flexible; I can help them find and vet different options, and connect them with resources. And I try to encourage creativity: sometimes the solution to sharing hard-to-share data is surprisingly analogue, or artistic, or ephemeral. Hard-to-share data does not get shared if we make it too daunting for the researcher to attempt, and it does not get used if we don't meet the intended audience halfway."

3 View Sabina Leonelli's ORCID profile here: <https://orcid.org/0000-0002-7815-6609>



# Case studies of 'hard-to-share' data in the Social Sciences and Humanities

In this section, we provide a series of case studies of researchers and research projects that have considered and attempted to address the bottlenecks involved in managing and sharing 'hard-to-share' data in various disciplines across the SSH. Detailed guidance on how to approach making hard-to-share data FAIR is available in a variety of educational and training resources, including the ones created by our project. However, we also collected case studies, because they provide real-life examples that may be inspiring and useful, providing realistic examples of how researchers have tackled the challenges at hand.

The examples below represent a broad group of projects that include a wide range of data types and different bottlenecks to sharing data. They cover hard-to-share data in corporate crime research in criminology; topically-sensitive qualitative interviews in psychology; and location-sensitive data in archeology. The source of information for these case studies was presentations by

the researchers themselves, made during the first workshop of our training series, which was held in The Hague in March 2025.

These researchers range from a PhD candidate to a university distinguished professor, to a researcher-turned-data-professional. This demonstrates the range of individuals and research teams who are tackling these bottlenecks and who are able to share their challenges and solutions to others through these case studies:

## **Case Study 1:** Navigating Hard-to-Share Data in Corporate Crime Research

Insights from Longitudinal Research on  
Non-Compliance in the Financial Industry

### **Profile of the research:**

Frederike Oberheim is a PhD candidate at the Netherlands Institute for the Study of Crime

---

4 View Frederike Oberheim's ORCID profile here: <https://orcid.org/0009-0008-2286-3516>



and Law Enforcement and is also affiliated with Leiden University.<sup>4</sup> Her research aims to provide an understanding of the mechanisms behind corporate non-compliance to legislation in the financial industry, for instance, of money laundering. The research seeks to find patterns of corporate risk behaviour over time and to relate those patterns to a) inspectorate behaviour (e.g. to what extent inspections are being carried out and how measures are being enforced) and b) observed non-compliance.

Further investigations can be into the key characteristics of corporations with higher levels of corporate risk behaviour and/or the comparative impact of enforcement measures.

#### **About the data:**

The data collected in this research consists partly of publicly available information that can be found in annual reports and financial statements. However, what makes the data particularly hard to access, let alone share, is **regulatory data, such as companies' survey responses on their own risk behaviour, risk profiles created by the regulatory authorities, reports on inspections, and correspondence on enforcement actions.**

This data is provided by De Nederlandsche Bank (DNB) and the Autoriteit Financiële Markten (AFM). The data access is arranged directly with those authorities, in line with their procedures and legal responsibilities. Access is granted to authorised researchers under formal agreements and with appropriate safeguards.

All analyses are conducted using secure environments<sup>5</sup> and established confidentiality protocols. Methods include descriptive analysis as well as exploratory factor analysis, multiple regression analysis, and longitudinal analysis, such as multi-group multi-trajectory modelling, latent class approaches, and growth curve modelling, applied as appropriate to the research questions and data permissions.

#### **The challenges with data sharing:**

Given the sensitivity of supervisory data, partner authorities **carry out statutory checks prior to any dissemination to ensure that confidentiality requirements are met and that organisations are not identifiable.**

These checks are a normal part of working with protected data and are designed to uphold legal obligations around privacy and supervisory secrecy. The research team **plans publications in a way that accommodates these processes** while maintaining academic independence, e.g., by agreeing on timelines for reviews, documenting any constraints transparently, and clearly separating methodological choices (which remain with the researchers) from disclosure control procedures (which remain with the data providers).

Working with protected data can involve additional steps, such as **coordinating access logistics, meeting information-security standards, and ensuring that materials not originally collected for research are used proportionately and within scope.** These

<sup>5</sup> For more information about one type of secure environment, see the section below on SANE.



steps help safeguard both research subjects and data providers and support the long-term feasibility of research with sensitive sources.

### **Addressing the challenges:**

In the presentation, Frederike Oberheim outlined concrete practices that have supported the research progress so far: (1) Early planning and clear agreements, including initiating access and data-management procedures at the project start, as well as defining roles, timelines, and disclosure review steps in formalised agreements; (2) Relationship-building and ongoing communication, such as maintaining regular dialogue with data providers about research aims, timelines, and any evolving needs; (3) Secure infrastructure including approved, secure storage and analysis environments and document controls in a Data Management Plan; and (4) Transparency about what can be shared, e.g., focusing on

sharing study designs, code, and synthetic or aggregated outputs where possible, and being explicit about limits that protect confidentiality.

## **Case Study 2:** Sensitive Qualitative Interviews

### **Profile of the research:**

Dr Rebecca Campbell is a community psychologist who conducts community-based field research and participatory action research that focuses on sexual assault survivors. The community focus was in Detroit, Michigan in the USA. Dr Campbell works at Michigan University.<sup>6</sup>

### **About the data:**

The data collected in this research is **qualitative data, namely interviews with sexual assault survivors that recount the assault and how police initially responded**

‘Even when data sharing is constrained, **prioritising openness when feasible** – regarding methods, code, and meta documentation – support reproducibility while respecting legal and ethical boundaries.’

### **Frederike Oberheim**

researcher

<sup>6</sup> View Rebecca Campbell's ORCID profile here: <https://orcid.org/0000-0003-0442-9835>.

**and treated them.** The interviewees describe the prosecution of their cases. The data are hard-to share because they contain indirect and direct personal data of the interview subjects. However, the data are *especially* sensitive due to topical sensitivities: the data has the topic of sexual assault. Thus there are particularly acute risks of re-identification of research participants because they recount sexual assault and their experiences of the criminal justice system.

### **The challenges with data sharing:**

Rebecca Campbell is open minded about open science and believes in transparency as a value and practice. She has shared data in data archives, but felt no immediacy or urgency in doing this until the Office on Violence Against Women in 2018 indicated that it would require datasets funded by them (including this research) to be submitted for archiving with the National Archive of Criminal Justice Data (NACJD).

This created a concern about how to create a safe, trauma-informed experience, **how to protect their privacy and confidentiality and, importantly, how to prevent the re-identification of their data.**

### **Addressing the challenges:**

This project tackles the challenges of sharing this topically sensitive data as a multi-phase process, a careful and diligent approach to counter the risk of **inadvertent disclosure of personal information.** Before the data collection began, the participants were asked for informed consent for data archiving: a crucial step in data management planning to make future data sharing possible. This also gave the participants an opportunity to

decide what they chose to disclose during the interview accordingly. Phase 1 was to develop a process to recognise potentially identifiable data, evaluating each data point for risk of re-identification.

Phase 2 involved the **remediation or redaction of this potentially identifiable data.** Remediation in this project is used for editing activities; redaction is used for removal or replacement with a summary. Teams were formed including those with and without deep knowledge of interviews, and these people read, re-read, and re-read again the transcripts and tagged information that was at risk of reidentification and a plan was made for each. This process involved multiple rounds of review and discussion and multiple checks that the plans were correctly implemented.

All of these approaches involve varying levels of **trade-offs in the usability of the data,** and also the researcher decided to redact information rather than to risk 're-writing' the survivor's story by remediating it. In some cases, descriptions of experiences with the legal system were retained because they were not deemed to be identifiable: many survivors had the same experience. Rebecca Campbell notes that it is **important to consult with the participants themselves:** 'for this study, we asked sexual-assault survivors at the end of the interview what information they wanted removed from their transcript before archiving. Survivors rarely asked for specific redactions, likely because we had already promised them in the informed-consent process that we would remove "other identifiable information" before archiving' (Campbell, 2023).

The final phase was to provide agency staff with a set of remediated transcripts asking, 'could they re-identify the survivor?'. The answer was no, demonstrating that the process was successful.

### Case Study 3: Sharing location data: Sensitive data in archaeology

#### Profile of the research:

Archaeological datasets often include spatial data: it is essential to know where an archaeological site is in the landscape, where a building or other feature is within a site, and in what vertical layer a feature or artefact was found. However, **location data of archaeological sites, especially, can be sensitive** and this information could be used by looters. Pascal Flohr (Leiden University and University of Oxford) and Adam Benfer (Leiden University) have both been working on projects where such archaeological location data are being collected.

#### About the data:

While archaeological datasets contain all sorts of qualitative and quantitative data, the location data themselves are in general simple N and E coordinates. Textual descriptions, sketches, and photos certainly also give clues on the location of a site, but less directly. Sensitive location data are not limited to archaeology, but also occur for example in ecology (endangered species), geology (e.g. location of precious metals), and various social sciences (e.g. location tracking of people).

#### The challenges with data sharing:

The challenge is to **find the right balance between having a useful, reusable dataset (i.e. with location data) but at the same time not risking the data being used by looters**. The question is how to decide on what to share and to what extent. It is also important to take into account legal restrictions (in some countries location data of heritage places cannot be shared at all, in others it is required to get permission from the authorities) and ethical considerations (whose data are they and who should (co)decide?).

#### Addressing the challenges:

Firstly, the researchers used a decision framework from ecology to help them determine to what extent to share location data (Chapman, 2020). It **sets out criteria for determining the sensitivity of taxa (or archaeological sites) and data in ecology (now applied to archaeology)**:

- Risk of harm: is the taxon (site) likely subject to harmful human activity?
- Impact of harm: how sensitive is the taxon (site) to the harmful human activity?
- Sensitivity of the data: will the release of the data increase the harm?

Secondly, the researchers **gave several options for the partial obfuscation of location data**, since it is almost never necessary or useful to completely leave out the location data. The following list is certainly not exhaustive, nor does it indicate a 'right' answer or preference, as it will depend on the situation which option works best:

- Decreasing the precision or resolution of the coordinates, for example by 0.1 degree.
- Using hexagons around the location (the Discrete Global Grid System) (Caspari et al., 2024)
- Obscuring the location by adding a buffer around the point or polygon
- Replacing the coordinates by the name of an administrative unit
- Coordinate modification by randomisation.

Finally, part of the solution can be to **provide different levels of access to the data**, for example having the detailed location visible for fellow researchers, but only the general area for public access.





# A solution for sharing hard-to-share data in the SSH: Secure ANalysis Environment (SANE) for inspecting and analysing data without downloads

One important challenge in data sharing that is often demanded by the data owner or provider, is the security of the infrastructure itself where the data will be shared. Data providers may be willing to share sensitive data, but not transfer the data to the researcher. The risk of data propagation or leakage is deemed too high in many cases. Sharing via emails, popular/commercial data transfer services, and data storage solutions (including personal devices) poses a huge security risk, often leading to unwillingness to share data altogether.

One effective way to enable secure data sharing, and particularly the reuse of sensitive data for new research, is through Trusted Research Environments (TREs). SANE (Secure

ANalysis Environment) is SURF's TRE, providing a controlled compute and analysis environment in which data can be provided for inspection and analysis without being downloadable, thereby leaving the data provider in full control.<sup>7</sup> Following the Five Safes framework, it is a proven solution to share sensitive data whilst allowing the researcher freedom in performing their analysis (see UK Data Service, n.d.).

SANE is a discipline-agnostic SURF service (ISO 27001 certified) running on SURF Research Cloud. SANE is currently used for studying sensitive data including privacy-sensitive data. During a hands-on demonstration in our training workshop, participants discussed how useful SANE is for other types of sensitive data. Availability of

<sup>7</sup> To learn more about SANE, see:

<https://www.surf.nl/en/themes/research-infrastructure/sane-secure-environment-for-analysing-sensitive-data>

such a certified, highly secure environment, hosted by a neutral organisation, can boost data sharing while providing researchers freedom to extract value from the data.

## **An example of SANE in action: the FIRMBACKBONE infrastructure**

In this section, we provide a case study of SANE's use in research, focusing on FIRMBACKBONE, an infrastructure that uses commercially sensitive data on all Dutch firms.<sup>8</sup> FIRMBACKBONE uses an access protocol that makes data accessible through SANE. The infrastructure is a part of ODISSEI, the Open Data Infrastructure for Social Science and Economic Innovations.

### **Corporate register as the backbone**

The FIRMBACKBONE data-infrastructure is built around the corporate register of the Netherlands maintained by the Dutch Chamber of Commerce (KVK).<sup>9</sup> Each firm and its establishments obtain a set of unique identifiers, making it easy to track these organisational entities over time. This structure has allowed the database to be supplemented with additional information. The value of FIRMBACKBONE lies in the possibility to combine various data contributions based on this corporate register. The data-infrastructure also provides employment data per establishment and financial statements for each firm. FIRMBACKBONE generates extra information from firm websites through web-scraping, based on data from the SIDN. Additional datasets from researchers have

been added using the corporate register identifiers, e.g. firm-specific interviews and environmental data.

### **Valuable researcher data**

Researchers in economics, sociology, innovation studies and related fields can access the complete corporate register. Meaning, all firms and their establishments are included in the dataset. The data has a longitudinal character enabling rigorous analysis of firm entry and exit dynamics, sectoral evolution, regional clustering and policy impacts. Scholars are able to investigate labour-market mobility, innovation diffusion, corporate-network structures and also language development patterns amongst firms with high granularity, supporting both descriptive mapping and causal inference.

### **Dealing with sensitive data using ODISSEI**

The FIRMBACKBONE data-infrastructure contains sensitive data due to its combination of the corporate register of the Netherlands with other information (financial statements, employment information, web-scraped data, and researcher contributions). This combined data is commercially very valuable. Due to this sensitive nature, data cannot be shared without due diligence and absolute guarantees of security.

FIRMBACKBONE was able to share the sensitive data based on ODISSEI's technical infrastructure for secure data sharing through SANE. While FIRMBACKBONE focusses on the data acquisition, data quality, researcher

<sup>8</sup> See <https://firmbackbone.nl/>

<sup>9</sup> See <https://www.kvk.nl/en/>

documentation and technical knowledge of the rich database, ODISSEI organises and maintains the infrastructure that makes access and publication of these data possible. This consortium model demonstrated both scholarly merit and rigorous safeguards which has been highly appreciated by data providers KVK, LISA and SIDN.

### **Safeguards adopted to assure data providers of sensitive data**

FIRMBACKBONE understood early in the process of acquiring data from data providers that the safety of the data is a key concern. Without a reliable platform data providers would not be convinced that their data would be in safe hands and would not go along with handing over a copy of their data. For this reason, FIRMBACKBONE, in conjunction with ODISSEI, adopted the following safeguards:

- **Five Safes framework:** Ensuring safe people, safe projects, safe settings, safe data and safe outputs
- **ISO 27001 certification:** Independent audit and penetration testing of the SANE environment
- **Audit trails:** Comprehensive logging of every data access, query execution and export operation
- **Research-only access agreements:** Each request evaluates the research purpose, and have researchers sign binding data-use agreements before access is granted.

### **Researcher requirements to work with the data**

To make the data-infrastructure attractive for researchers to work with requires information on the data and an environment in which they can:

- **Learn about the data:** Extensive meta-data on the availability and an evaluation of data quality must be combined with clear instructions on data access requirements.
- **Work with the data:** Inspect, query and manipulate all observations and variables to explore hypotheses and refine analytical code.
- **Have guaranteed reproducibility:** Work with static biannual snapshots of the FIRMBACKBONE database to ensure consistent, peer-reviewable results.
- **Use flexible and on-demand compute:** Provision script-based environments that scale to the computational complexity of their analyses.
- **Linkage Keys:** the use of unique firm information allows joining auxiliary datasets.

### **The role of FIRMBACKBONE and the Five Safes**

FIRMBACKBONE serves as the operational custodian of the combined datasets.

It manages and ingests biannual data of the raw corporate register dataset from the KVK and other data providers. An extensive and rigorous procedure is in place to verify and combine these data. FIRMBACKBONE enforces the Five Safes framework.

- **Safe data:** achieved by pseudonymisation before it gets published and made available to researchers.
- **Safe people:** providing access to researchers is based on strict (maximally 6 months of access) and clear procedures (described in the meta-data portal).
- **Safe project:** based on reason of the access request.



- **Safe settings:** the usage of the SANE infrastructure with secure and on-demand compute environments.
- **Safe output:** output checks are conducted for all exported material to prevent identifiable data disclosure (e.g., through low cell counts), and executes researcher due-diligence and agreement processes.

With this framework, FIRMBACKBONE ensures that every access request, computation and export upholds compliance in line with the consortium's accountability standards.

### **More on SANE's potential use in different research projects**

In this case study, FIRMBACKBONE achieved access to a proliferating amount of sensitive and valuable data from multiple data

providers, including the renowned corporate register data from KVK. For a large part these providers recognised the safe circumstances in which access to the data is provided, and continue to support this initiative. The same SANE environment can also be used by other case studies discussed here, where an individual researcher, project or an organisation can play a role of data provider. They choose who to collaborate with and provide data to, while always being in charge of their data.



# Final thoughts

The idea of data being ‘hard-to-share’ by itself is not one that we can define in precise terms, nor did we aim to do so, though some participants in our training workshops would perhaps have liked to be given such clarity, as it facilitates easy decision making.

A nice way to illustrate this, may be by using the analogy of a traffic light: a green light for data that are easy to share or a red light for data that cannot be shared at all. ‘Hard to share’ will be orange, asking us to watch out, look around at what is happening, oversee the circumstances, and then to make a well balanced, reasoned decision, knowing that we still have to be careful. Sometimes, on closer inspection, the orange may be closer to red.

As we prepared our training workshop, it became clear that what makes a dataset hard to share may be hard to explain. The fact that qualitative data, often found in SSH research, are diverse in nature, results in complications with sharing, which can have multiple reasons, and those reasons themselves can also be hard to trace. In some of the case studies above, we have showed that a dataset may well contain some personal data, but that the GDPR was just not the *main* restriction for being cautious about sharing. Nor does a dataset have to be sensitive in order to be hard to share: it could also ‘just’ contain items on which a rightsholder places restrictions.

We want to briefly return to highlight two of the case studies that we have presented above. Rebecca Campbell’s research project

was insightful. We witness the significant effort that Dr Campbell has made to be able to share, in meaningful way, some very sensitive data with the purpose of doing justice to rape victims, who want their stories to be told, to be heard. It would have been much easier for the research team to have kept the data to themselves. However, with consideration, she has carefully navigated this bottleneck.

Another source of inspiration was the presentation by PhD candidate Frederike Oberheim who closed with some lessons learned:

- Focus on what you can share
- Be transparent about your limitations

In this guide, we have touched on research data in a wide variety of SSH domains, ranging from archaeology to economics, and we have presented SANE: a secure environment for analysing sensitive data.

All in all, it seems only logical that the complex nature of qualitative datasets can lead to a complex discussion. We trust that our audience was sufficiently familiar with this field of research to appreciate the examples, the opportunities for in depth discussion, rather than being provided with ‘yes or no’ answers. We see this as a topic that will benefit from further exploration, richer training and educational opportunities and resources, and further careful attention of researchers and research supporters.





# Training materials from the three-part workshop

This guidebook is based on a three-part series of training workshops. The project team created detailed session plans for each workshop, as well as slide decks for the various presentations, and the plans for all of our group exercises. Even though the workshops were attended by, in total, 95 researchers and data professionals, it was not possible for everyone who was interested in the topics to attend these one-off in-person sessions.

So, having put considerable effort into creating these innovative training materials, the project team deemed that it was important to share them more widely, making them available for reuse by other trainers in the future. To draw your attention to these, and to invite you to reuse them under their CC-BY licence, we have provided links to them in Zenodo below:

Flohr, P., van den Berk, M., Roodhof, A., Verheijen, J., Bruil, M., Thorpe, D.E. (2025), 'Workshop - Sharing Field Notes'. Zenodo.

Available at: <https://doi.org/10.5281/zenodo.15600311>

Lushaj, B., Gelens, T., Magraw, J.-Y., Mos, A., Baloum, R.-C., Hati Gitundu, (Beatrice) BH. (2025), 'Workshop on The Ethics of Sharing Fieldwork Data and the CARE Principles'. Zenodo.

Available at: <https://doi.org/10.5281/zenodo.15629394>

Thorpe, D.E., van den Berk, M., Flohr, P., van der Meer, L., Hesam, A., Campbell, R., Oberheim, F. (2025), 'Workshop on Hard to Share Data in the Social Sciences and Humanities and using the Secure ANalysis Environment (SANE)'. Zenodo.

Available at: <https://doi.org/10.5281/zenodo.15302953>





# References

## C

Carroll, S.R., Garba, I., Figueroa-Rodríguez, O.L., Holbrook, J., Lovett, R., Materechera, S., Parsons, M., Raseroka, K., Rodriguez-Lonebear, D., Rowe, R., Sara, R., Walker, J.D., Anderson, J., Hudson, M. (2020), 'The CARE Principles for Indigenous Data Governance', *Data Science Journal*, 19. <https://doi.org/10.5334/dsj-2020-043>

CESSDA Training Team (2017 - 2022), 'Anonymisation, CESSDA Data Management Expert Guide'. Available at: <https://dmeg.cessda.eu/Data-Management-Expert-Guide/5.-Protect/Anonymisation> (Accessed: 21 October 2025)

Campbell, R., Javorka, M., Engleton, J., Fishwick, K., Gregory, K., Goodman-Williams, R. (2023), 'Open-Science Guidance for Qualitative Research: An Empirically Validated Approach for De-Identifying Sensitive Narrative Data'. *Advances in Methods and Practices in Psychological Science*, 6(4). <https://doi.org/10.1177/25152459231205832>

Caspari, G., dos Santos Manuel, J., Gago-Silva, A. et al. (2024), 'Employing discrete global grid systems for reproducible data obfuscation', *Sci Data* 11(509). <https://doi.org/10.1038/s41597-024-03354-5>

Chapman, A.D. (2020), 'Current Best Practices for Generalizing Sensitive Species Occurrence Data', Copenhagen: GBIF Secretariat. <https://doi.org/10.15468/doc-5jp4-5g10>

## F

Flohr, P. (2024), 'Sharing location data: Sensitive data in archaeology', Zenodo. <https://doi.org/10.5281/zenodo.12706680>

Flohr, P., van den Berk, M., Roodhof, A., Verheijen, J., Bruil, M., & Thorpe, D. E. (2025), 'Workshop - Sharing Field Notes', Zenodo. <https://doi.org/10.5281/zenodo.15600311>

Flohr, P., Roodhof, A., 'Sharing Field Notes', Zenodo. Reserved DOI: <https://doi.org/10.5281/zenodo.17588823>

## G

General Data Protection Regulation (GDPR) (n.d) *General Data Protection Regulation (GDPR)*. Available at: <https://gdpr-info.eu/> (Accessed 21 October 2025)



Government of the Netherlands (n.d) *What is knowledge security? | National Contact Point for Knowledge Security*. Available at: <https://english.loketkennisveiligheid.nl/knowledge-security> (Accessed 21 October 2025)

Grau, N., Thorpe, D., L'Hours, H., David, R., Flohr, P., Krüsselmann, K. (2024), 'Where do I start with FAIRification of sensitive data?', Zenodo. <https://doi.org/10.5281/zenodo.12733678>

## I

International Red Cross (1965), 'Proclamation of the fundamental principles of the red cross', *International Review of the Red Cross* 56, pp. 573-576

## J

Jaber, S. (2022) *Webinar "FAIR for Sensitive Data": A Short Recap. GO FAIR*. Available at: <https://www.go-fair.org/2022/04/13/webinar-fair-for-sensitive-data-a-short-recap/> (Accessed 21 October 2025).

## K

Kleef, S. van, Ploeg, J.L. van der P., Moester, M., Hoogen, H. van den, Jansen, E., Meer, T. van der, Romero Pastrana, F., Scholten, J., Spek, L. van der, Verheul, I., (2019), 'Dealing with pseudonymization and key files in small-scale research. A few basic steps', Zenodo. <https://doi.org/10.5281/ZENODO.3571046>

## L

Lushaj, B., Gelens, T., Magraw, J.-Y., Mos, A., Baloum, R.-C., Hati Gitundu, (Beatrice) BH (2025), 'Workshop on The Ethics of Sharing Fieldwork Data and the CARE Principles', Zenodo. <https://doi.org/10.5281/zenodo.15629394>

Lushaj, Bora et al. 'The Ethics of Sharing Fieldwork Data and the CARE Principles', Zenodo. Reserved DOI: <https://doi.org/10.5281/zenodo.17588886>

## R

Ravenzwaaij, D. van, de Jong, M., Hoekstra, R., Scheibe, S., Span, M.M., Wessel, I., Heininga, V.E. (2025), 'De-Identification When Making Data Sets Findable, Accessible, Interoperable, and Reusable (FAIR): Two Worked Examples From the Behavioral and Social Sciences', *Advances in Methods and Practices in Psychological Science* 8. <https://doi.org/10.1177/25152459251336130>

## T

Thorpe, D., Braukmann, R. (2024), 'Balancing privacy and reusability: The why, what, and how of de-identifying research data', Zenodo. <https://doi.org/10.5281/zenodo.13772647>



Thorpe, D.E., van den Berk, M., Flohr, P., van der Meer, L., Hesam, A., Campbell, R., Oberheim, F. (2025). Workshop on Hard to Share Data in the Social Sciences and Humanities and using the Secure ANalysis Environment (SANE), Zenodo. <https://doi.org/10.5281/zenodo.15302953>

## V

Verburg, M., Braukmann, R., Mahabier, W. (2023), 'Making Qualitative Data Reusable - A Short Guidebook For Researchers And Data Stewards Working With Qualitative Data', Zenodo. <https://doi.org/10.5281/zenodo.8160880>

## U

UK Data Service (n.d) *What is the Five Safes framework?* Available at: <https://ukdataservice.ac.uk/help/secure-lab/what-is-the-five-safes-framework/> (Accessed 21 October 2025)







This publication is part of the project 'Beyond personal data: a new initiative to support early-career researchers with hard-to-share data' with file number ICT.TDCC.001.002, which is (partly) financed by the Dutch Research Council (NWO) via the Thematic Digital Competence Centre Social Sciences & Humanities (TDCC-SSH).

