FORS

explore.understand.share.

Managing and Anonymizing Sensitive Data

Marieke Heers

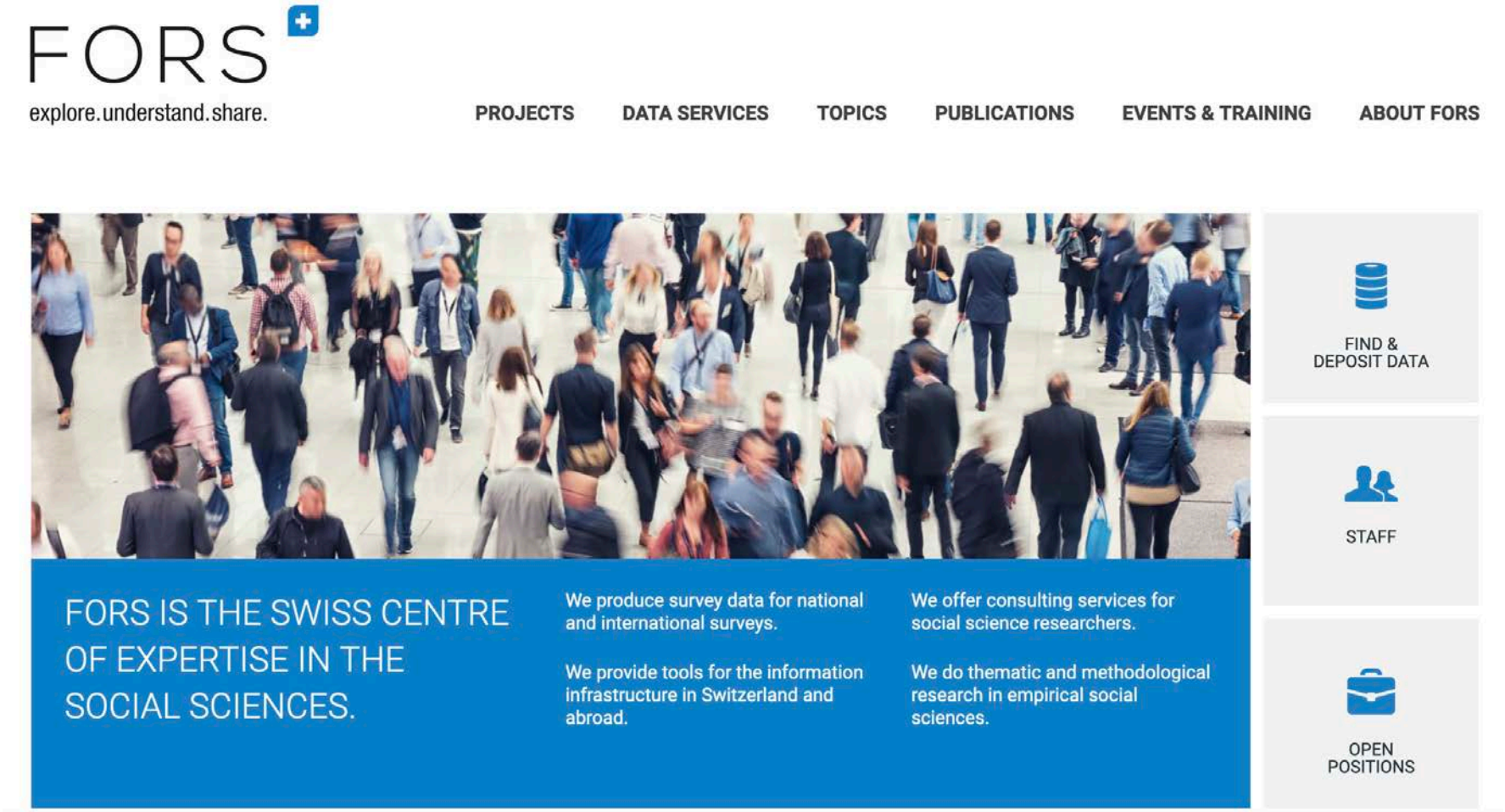
8 May 2023

Outline

1. FORS
2. Data management
3. Normative frameworks for data management
4. Personal and sensitive data
5. Anonymization of research data
6. Discussion and resources

FORS

Swiss Centre of Expertise in the Social Sciences
www.forscenter.ch



FORS

FORS – Swiss Centre of Expertise in the Social Sciences

- Research infrastructure of national scope intended for any institution or person active in the social sciences
- Mostly funded by the Swiss National Science Foundation (SNSF) and hosted by the University of Lausanne
- Founded in 2008

FORS

FORS COVID-19 SURVEYS →

METHODOLOGICAL RESEARCH →

POLITICAL PARTICIPATION AND PUBLIC OPINION →

VALUES AND ATTITUDES →

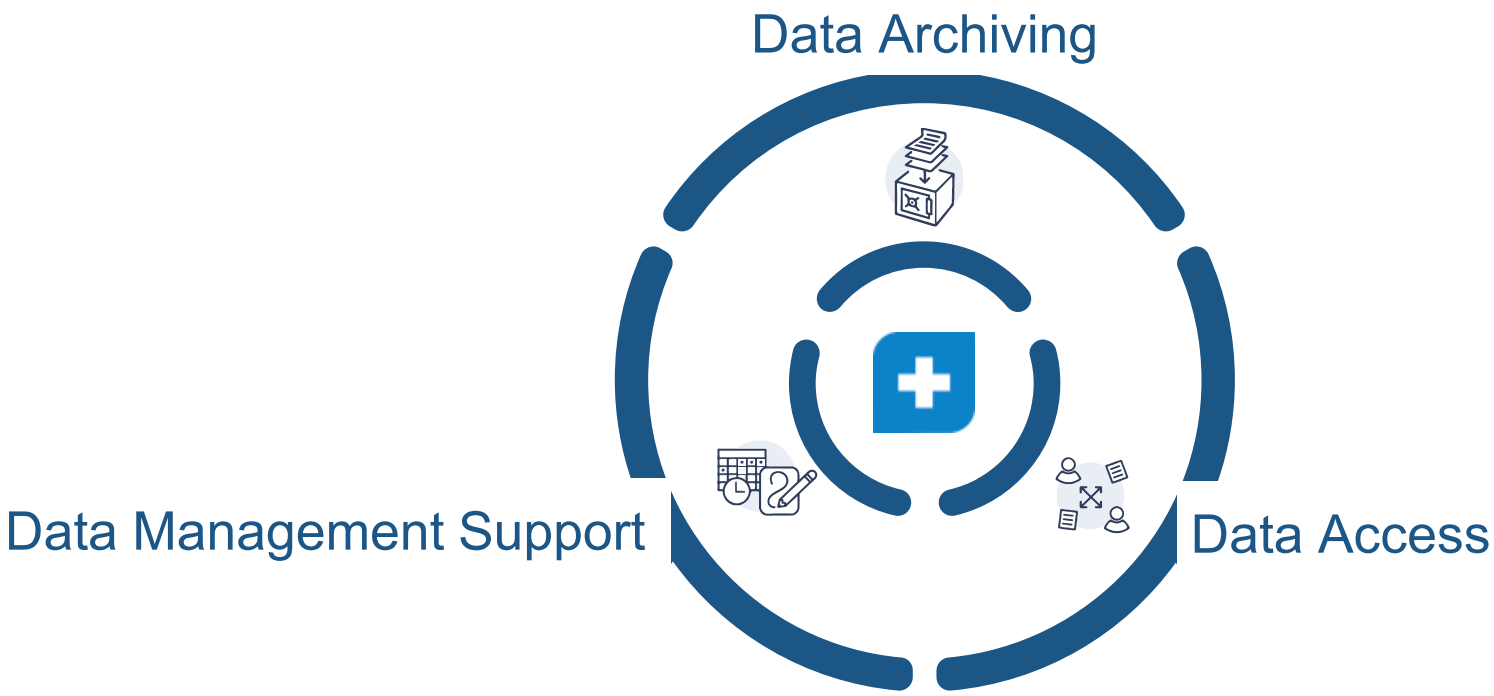
SOCIAL CHANGE →

LIFE COURSE →

WELLBEING →

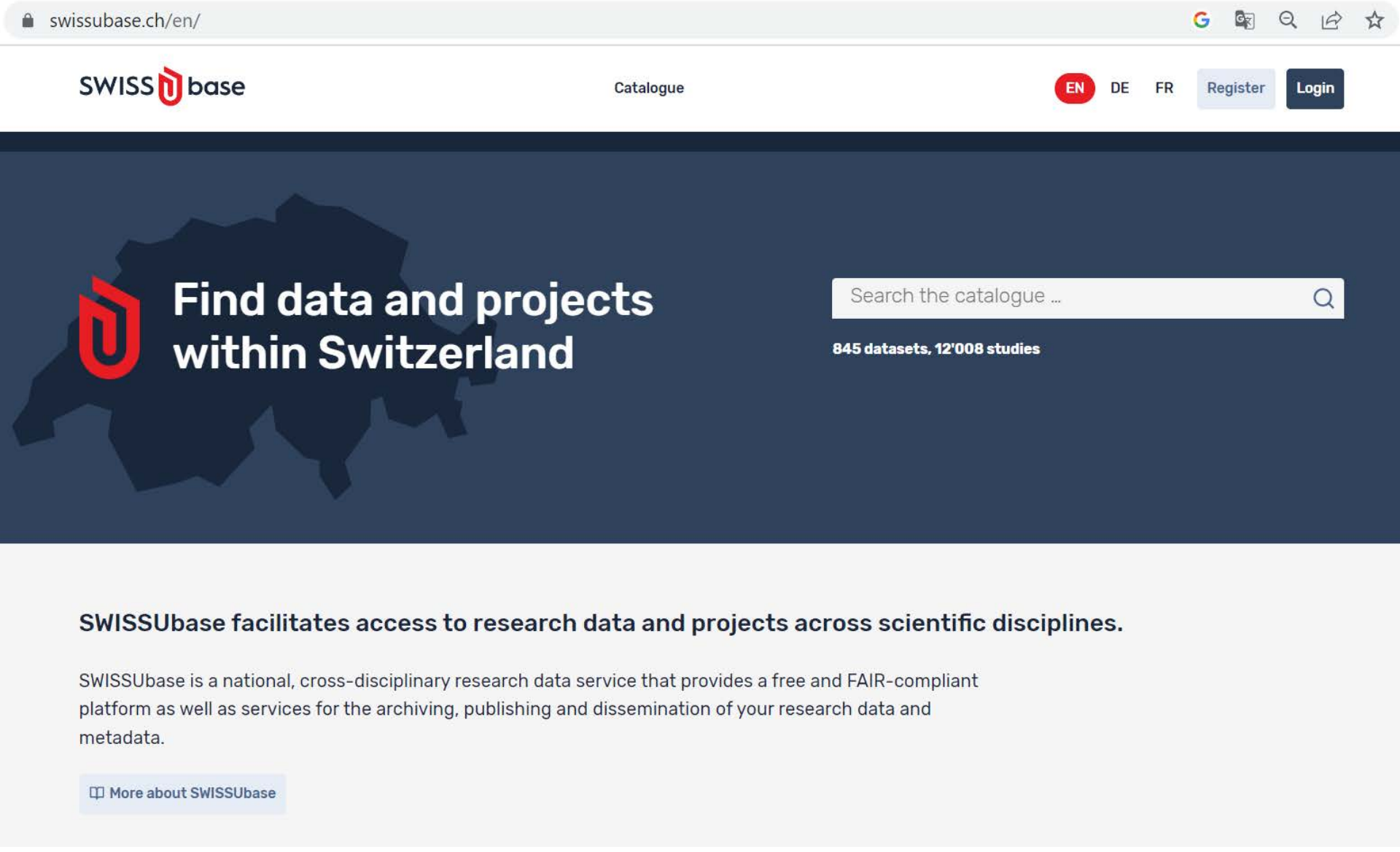
FORS

FORS Data Archive Services

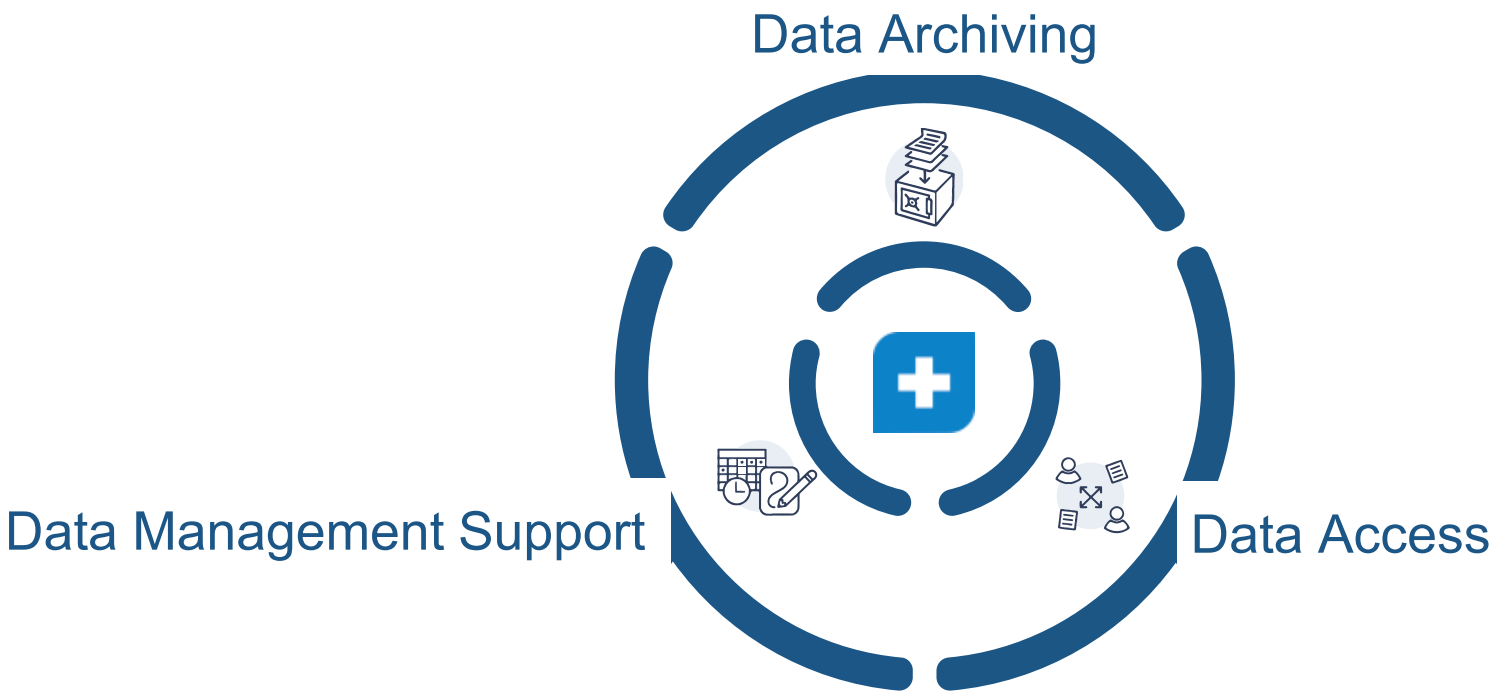


FORS

www.swissubase.ch



FORS Data Archive Services



FORS

Data management support at FORS

Alexandra Stam

Marieke Heers

Pablo Diaz

FORS

What is data management?

FORS

Definition: Research data management

All activities associated with data other than the direct collection and analysis

- Handling
- Organizing
- Documenting
- Enhancing
- Enabling sustainability and sharing

FORS

Research data lifecycle

Data management needs to be considered throughout your project:

1. At the start of your project, while planning how you intend to create, process and preserve your data
2. During the project, where you manage your data on a day-to-day basis
3. At the end of the project, to preserve the data and make them available whenever possible

Data management and sensitive data

- Sensitive data require specific considerations.

What normative framework applies to my data?

FORS

FORS GUIDES
to survey methods
and data management

Data protection: legal considerations for research in Switzerland

Pablo Diaz[1]

[1] FORS - UNIL

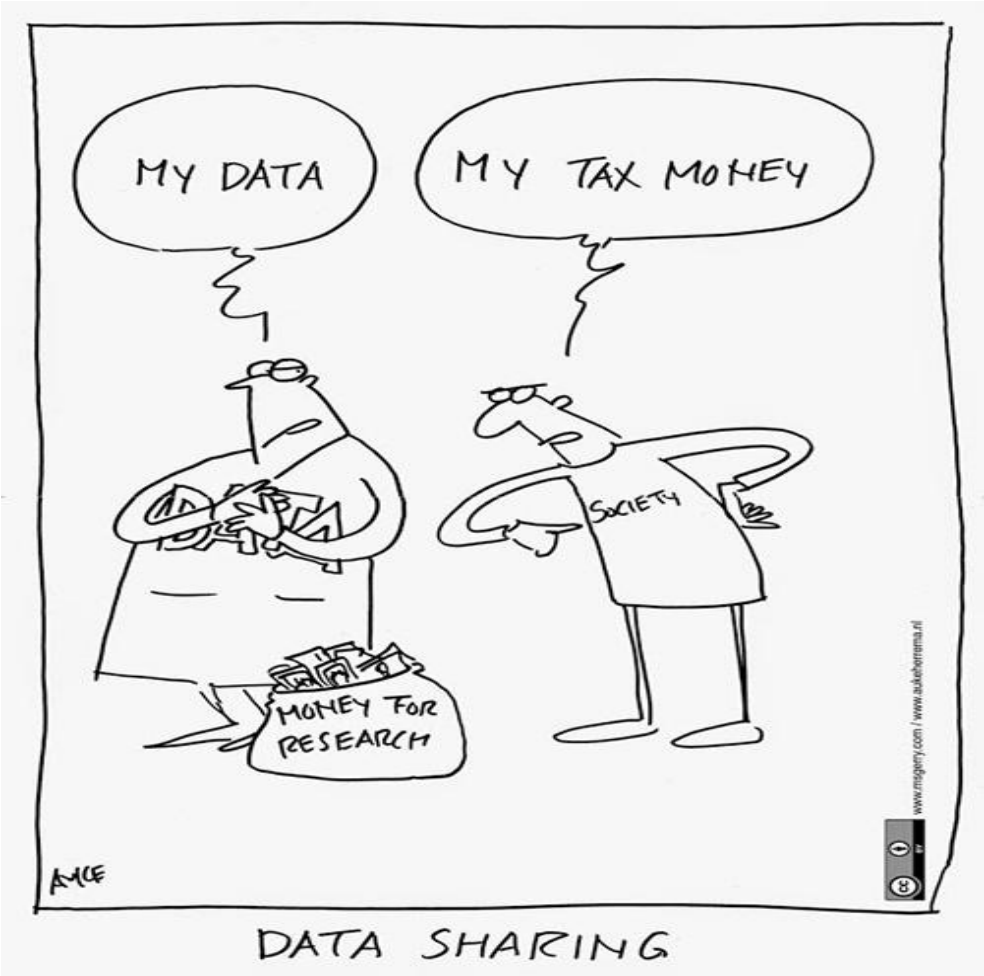
FORS Guide No. 17, Version 1.0

January 2022

FORS

Open data

- Data as public goods
- Reproducibility
- Secondary analysis
- …



FORS

Open data requirements

- From funders
- From research institutions
- From journals
- From peers

FORS

Data management and data protection

Personal and sensitive data cannot be managed without knowledge of existing laws. The processing of this type of data is governed by a legal rules related to **data protection**.

- Data protection aims to protect people's **privacy**.
- Privacy is generally referred to through the concept of **informational self-determination**.

Data protection

Informational self-determination is guaranteed by several **fundamental texts**:

- Universal Declaration of **Human Rights** (art. 12)
- European Convention on **Human Rights** (art. 8)
- Swiss Federal **Constitution** (art. 13)

Data protection

Swiss Federal **Constitution**:

Every person has the right to privacy in their private and family life and in their home, and in relation to their mail and telecommunications. (art. 13 al. 1)

Every person has the right to be protected against the misuse of their personal data. (art. 13 al. 2)

Data protection

- Anyone who processes personal data must comply with a number of rules and **laws**.
- In Switzerland, there are laws at two levels:
  - Federal (**FADP**)
  - Cantonal (**LPrD**, ...)
- In Europe there is the General Data Protection Regulation (**GDPR**)

Legal status

In Switzerland, the "legal status" of the data controller determines which law applies (federal or cantonal).

Private person/company	Federal body	Cantonal body
Federal laws	Federal laws EPFL, ETH, FORS, …	Cantonal laws UNIL, UNIGE, UNIBE, UNIZH, USI, HES, ...

If you carry out research as an employee of a university, you are subject to **cantonal laws** (since you work for the cantonal administration)

FORS

Legal status

Private person/company	Federal body	Cantonal body
Federal laws	Federal laws EPFL, ETH, FORS, …	Cantonal laws UNIL, UNIGE, UNIBE, UNIZH, USI, HES, ...

How about your research?

FORS

Am I processing personal data?

FORS

Personal data and data protection

- Knowing whether or not you are processing personal data is crucial.
- Processing personal data implies that you have to comply with **data protection** laws and regulations.

Are you processing personal data?

FORS

Personal data

"All information relating to an identified person" (Art. 3 lit. a FADP)

- **Very broad notion**: everything that can be related to a specific person is personal data!
- The most common: name, date of birth, home address, phone number, email, IP address, picture
- But also: opinions, original ideas, a style of writing, the way of walking, etc.

> *How you walk is just as unique as your fingerprint, your iris pattern and your voice!*

Before the pandemic, about 450,000 people walked through New York's Times Square every day! Now, what if I told you that each of these 450,000 individuals had a completely unique stroll through that space? Yes, every human being on Earth has their own singularly individual style of walking!

https://www.scienceabc.com/eyeopeners/is-your-walk-as-unique-as-your-fingerprint.html

Even Anonymous Coders Leave Fingerprints

Researchers have repeatedly shown that writing samples, even those in artificial languages, contain a unique fingerprint that's hard to hide.

RESEARCHERS WHO STUDY stylometry—the statistical analysis of linguistic style—have long known that writing is a unique, individualistic process. The vocabulary you select, your syntax, and your grammatical decisions leave behind a signature. Automated tools can now accurately identify the author of a forum post for example, as long as they have adequate training data to work with. But newer research shows that stylometry can also apply to *artificial* language samples, like code. Software developers, it turns out, leave behind a fingerprint as well.

Rachel Greenstadt, an associate professor of computer science at Drexel University, and Aylin Caliskan, Greenstadt's former PhD student and now an assistant professor at George Washington University, have found that code, like other forms of stylistic expression, are not anonymous. At the DefCon hacking conference Friday, the pair will present a number of studies they've conducted using machine learning techniques to de-anonymize the authors of code samples. Their work, some of which was funded by and conducted in collaboration with the United States Army Research Laboratory, could be useful in a plagiarism dispute, for instance, but also has privacy implications, especially for the thousands of developers who contribute open source code to the world.

Sensitive data

Personal data on: religious, ideological, political or trade-union related views or activities; health, the intimate sphere or the racial origin; social security measures; administrative or criminal proceedings and sanctions (Art3. lit. c FADP)

- The list provided by FADP is **exhaustive**
- In Switzerland, salary is not considered sensitive data
- Depending on the **context**, almost all data can be considered sensitive (name, photo, job, etc.)

Schweizer Haushalt-Panel
Panel suisse de ménages
Swiss Household Panel

580 | P20N50 | Original | USER | <= | all

Membre d'associations: Organisation ou groupe religieux	Vereinsmitglied: Religiöse Organisation oder Gruppe	Membro di associazioni: Organizzazione o gruppo religioso	Associational membership: Religious organisation or group
? Maintenant je vais vous lire une liste d'associations et d'organisations. Pouvez-vous me dire pour chacune d'elles si vous en êtes membre actif, membre passif ou pas membre?	? Jetzt lese ich Ihnen eine Liste von Organisationen und Vereinigungen vor. Könnten Sie mir für jede sagen, ob Sie Aktivmitglied, Passivmitglied oder nicht Mitglied sind.	? Ora Le leggerò una lista di associazioni e di organizzazioni. Può dirmi per ciascuna di queste se Lei ne è membro attivo, membro passivo o se non ne è membro?	? I will now read out a list of associations and organisations. Could you tell me for each of them whether you are an active member, a passive member or not a member?
! Membre passif = cotisation uniquement	! Passivmitglied = nur Beitrag	! Membro passivo = solo contributo	! Passive member = fee only
• Organisation ou groupe religieux.	• Religiöse Organisation oder Gruppe.	• Organizzazione o gruppo religioso	• Religious organisation or group.
-8 autre erreur	-8 anderer Fehler	-8 altro errore	-8 other error
-7 erreur de filtre	-7 Filterfehler	-7 errore di filtro	-7 filter error
-3 inapplicable	-3 trifft nicht zu	-3 inapplicabile	-3 inapplicable
-2 pas de réponse	-2 keine Antwort	-2 nessuna risposta	-2 no answer
-1 ne sait pas	-1 weiss nicht	-1 non sa	-1 does not know
1 Vous en êtes membre actif	1 Aktivmitglied sind	1 Lei ne è membro attivo	1 Active member
2 membre passif	2 Passivmitglied sind	2 membro passivo	2 Passive member
3 ou pas membre	3 oder nicht Mitglied	3 oppure non membro	3 Not a member

FORS

Do data protection laws apply to your research?

FORS

Am I processing personal data?

- In the social sciences, it is very difficult to have completely anonymous data.
- To be considered anonymous, **all information** that can be linked to an identifiable person must be **permanently destroyed.**
- A lot of information can potentially allow the identification of an individual.
- It is safer to assume that you are dealing with personal data.

FORS

Questions?

FORS

Anonymization of research data

FORS

Anonymization in the current research environment

- More and more data are produced
- New research fields, including new types of data (Big Data)
- Computational power allows for analysis of increasingly rich datasets
- Facilitated access to data by the community
- New analytical and data extraction tools

Anonymization in data management

- Anonymization is a key practice for protecting respondents and allowing data sharing.
- Anonymization needs to be understood in light of the legal and ethical requirements, but also in combination with other data management practices.

Do you face challenges regarding anonymization of your data?

FORS

Anonymization – A definition

- The notion of anonymization refers to the process by which the elements allowing the identification of a person are **definitively** deleted from a dataset, a document, an interview transcript, etc.
- As a result, an individual cannot be identified *without significant effort*.
- Represents a principal solution for complying with data protection requirements.
- This is **irreversible**!

Anonymization – A difficult promise

- Individuals are more unique than we might think!
- Crossing three simple variables, i.e. date of birth, postal code and gender, 63% of the US population can be identified (Golle, 2006).
- The collection of big data (via apps etc.) makes identification relatively easy.
- The ability to cross-reference research data with other datasets, information from social networks, blogs, or websites greatly facilitates (re)identification.
- Particularly relevant when working on small populations.

Anonymization – Requirements

- The data itself and all options of recreating the original data are eliminated completely
- The person can no longer be identified and the process is irreversible
- Fully anonymized data is no longer considered personal data
- The effort to identify the data subject is too big
  - in terms of know-how
  - in terms of cost
- It is very difficult to have fully anonymized data

Anonymization vs. pseudonymization

- The removal or replacement of identifiers with pseudonyms or codes.
- The data remain pseudonymous as long as the original identifying information exists.
- Enhances the security of the processed data by making the data subject identifiable instead of directly identified.
- Pseudonymized data remain personal data.

Anonymizing data: Factors to be considered

1. The nature and type of personal data to anonymize
2. The future users of the data and conditions of use
3. Balancing utility and data protection
4. Risk management
5. What was promised to respondents

1. The nature and type of personal data to anonymize

- Sensitivity
- Sampling
- Duration
- Data from other sources

2. Future users of the data and conditions of use

- Public release or restricted access
- Likely expertise of users
- Access conditions (e.g., with prior approval, with user contract)

3. Balancing utility and data protection

- Increased protection implies decreased utility
- Expected analyses
- Variable level assessment
- What can be sacrificed?

4. Risk management

- Motivations for an attack
- Consequences of a disclosure
- Disclosure without malicious intent
- Possibilities to link other data or knowledge

There is no risk-free scenario. The goal is to identify the acceptable level of risk.

5. Promises to respondents

What was promised to respondents in a consent form is ethically and legally binding.

Setting up an anonymization strategy

- What types of direct or indirect identifiers do my materials contain? Are there rare/unique information in the data?
- What combinations of variables can allow identification of an individual?
- Can information from other sources be linked to the data making identification possible? (social networks, blogs, etc.).
- What characteristics of the data do I want to retain and which ones can be “sacrificed”?

Direct and indirect identifiers

- Direct identifiers alone are sufficient to identify people (e.g., name, social security number)
- Strong indirect identifiers allow fairly easy identification (e.g., home address, telephone number)
- Weak indirect identifiers allow identification through *combinations* of variables

Indirect identifiers: Socio-demographic variables

- Gender
- Age (DOB, MOB, YOB)
- Location (municipality, canton, main region, linguistic region)

- Civil status
- Nationality
- …

FORS

Indirect identifiers: Geographical variables

Postal code level vs. canton/linguistic region

Basic approach

- Removal of direct and strong indirect identifiers
- Assessment of weak indirect identifiers and appropriate techniques
- Starting with a categorization of variables

Anonymization techniques

- Variable suppression
- Record suppression
- Character masking
- Pseudonymization
- Generalization
- Swapping
- Data perturbation

Variable suppression

- Removal of an entire variable
- Extreme loss of information, so should be last resort
- First technique to apply
- Often used with sensitive open-ended questions

Record suppression

- Removal of an entire record that cannot easily be anonymized (e.g., an exceptional and easily identifiable individual)
- First assess whether other techniques might handle the problem (e.g. generalization)

What does this imply for your sample?

Character masking

- Change of the characters of a data value, using a constant symbol (e.g. “*” or “x”)
- Partial hiding within a string
- Replace a fixed or variable number of characters

Example:

079 259 67 00 -> xxx xxx 67 00

078 452 83 14 -> xxx xxx 83 14

FORS

Pseudonymization

- Replace identifying information with made-up values
- For cases where values must be uniquely distinguished
- Made-up values must be arbitrary and unique
- Can be reversible or irreversible
- Often used to link individuals across datasets

Generalization

- Reduction of precision of a variable
- Create discrete categories from a continuous variable
- Combine values into broader categories
  - e.g. age, professions, income, …

Swapping

- Rearrange data across records such that the individual variable values are still represented in the dataset
- Only to be used when analysis is on aggregate level, i.e., where there is no need to examine relationships between variables

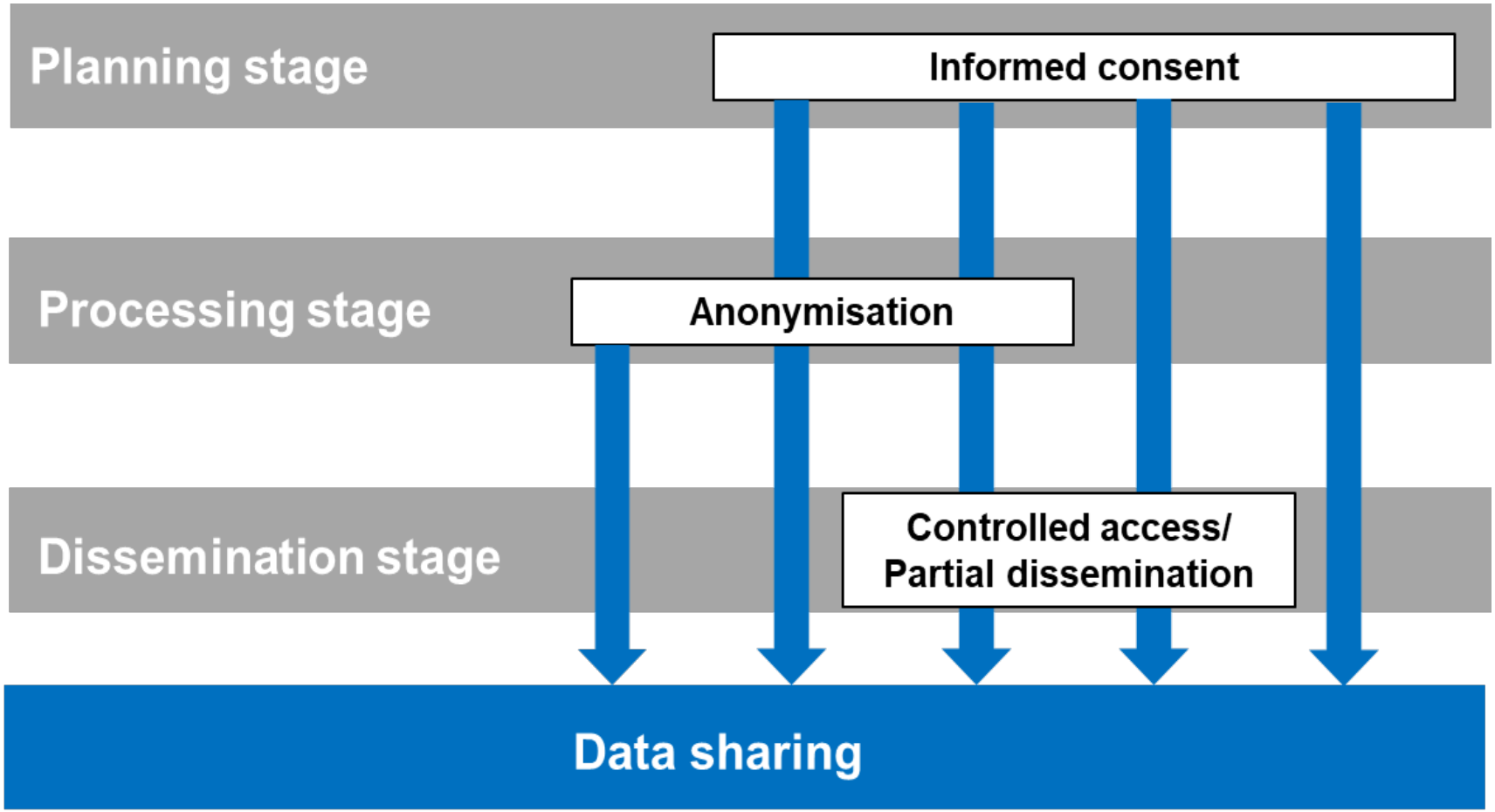
Data perturbation

- Modification of values to be slightly different
- Where small changes of value do not significantly affect analysis and accuracy
- Examples include base-x rounding and adding random noise

Some additional points to consider

- Minimize – ask only for what you need
- Anonymization should be considered together with consent agreements and access restrictions
- Regulating/restricting user access may offer a better solution than anonymizing
- Data that need anonymization should be avoided in data collection
- Direct identifiers should be removed, masked or changed
- A maximum of information should be maintained
- Unedited versions of data should be retained for preservation
- Anonymization should be planned at the beginning of the research, not at the end

Planning for data sharing



FORS

Questions?

FORS

Discussion

FORS

Win the FORS Data Re-use Award 2023

FORS DATA RE-USE AWARD 2023

Submission deadline: 01.10.2023

Are you currently working with social science data you got from FORS? Or are you planning to do so in the next few months? Then consider submitting your paper to the FORS Data Re-use Award 2023 and get the chance to win up to 1'000CHF. The FORS Data Re-use Award 2023 will be given to young researchers who carry out excellent research that is based on secondary data.

https://forscenter.ch/fors-data-reuse-award-2023/

FORS

Resources

FORS

FORS GUIDES

to survey methods and data management

The FORS Guides offer support to researchers and students in the social sciences who intend to collect data, as well as to teachers at university level, who want to teach their students the basics of survey methods and data management. Written by experts from inside and outside of FORS, the FORS Guides are descriptive papers that summarise practical knowledge concerning survey methods and data management. They give a general overview without claiming to be exhaustive. Considering the Swiss context, the FORS Guides can be especially helpful for researchers working in Switzerland or with Swiss data. The FORS Guides are an ongoing publication series, additional guides will follow soon.

Data protection: legal considerations for research in Switzerland

Pablo Diaz [1]

[1] FORS - UNIL

FORS Guide No. 17, Version 1.0

January 2022

Data anonymisation: legal, ethical, and strategic considerations

Alexandra Stam [1] **and Brian Kleiner** [1]

[1] FORS

FORS Guide No. 11, Version 1.1 (last update January 2022)

June 2020

Qualitative data anonymisation: theoretical and practical considerations for anonymising interview transcripts

Alexandra Stam [1] **and Pablo Diaz** [2]

[1] FORS

[2] FORS-UNIL

FORS Guide No. 20, Version 1.0

March 2023

DATA MANAGEMENT

With the advent of the digital turn and the new requirements of Open Science, data management has become in recent years one of the major challenges facing the scientific community. At FORS, we follow an integrated, tailor-made approach based on concrete research practices. Our team is committed to developing tools to help researchers optimize the scientific quality of their research materials, which take into account both the specificities of their projects and current standards.

We aim to answer questions such as: How can I meet ethical and legal standards while respecting the epistemological and methodological codes of my discipline? How to respond to the growing demand for data openness without violating the new data protection rules? How do I store my data securely? Can I share everything? What will happen to my data after I retire?

You will find here useful resources that will help you to develop a strategy adapted to your needs, be it for data management planning or its implementation throughout a research project. Much more than an administrative procedure, data management is an opportunity to reflect on fundamental issues related to the production and processing of data, but also on the potential of data beyond the research projects for which they were initially collected.

Contact us!

- For support with data access and deposit:

  dataservice@fors.unil.ch
- marieke.heers@unil.ch

FORS

Thank you!

FORS

Additional resources

- CESSDA Data Management Expert Guide:

  https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert-Guide
- https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/Guide-to-anonymization_v1-(250118).pdf
- https://www.fsd.uta.fi/aineistonhallinta/en/anonymization-and-identifiers.html

FORS