FORS data management webinar series

Qualitative data Anonymisation

Dr. Alexandra Stam

November 22, 2022

FORS

FORS webinar series 2022

1. Informed consent / September 27th

2. Data documentation / October 11th

3. Quantitative data anonymisation/ November 1st

4. Qualitative data anonymisation

Outline

1. Introduction
2. Defining qualitative data anonymisation
3. Anonymisation as a process: towards a layered approach
4. Some common forms of resistance
5. Proceeding to data anonymisation
6. Top 10 recommendations

Introduction

FORS

Changing landscape

Open data: new requirements from funders, institutions and journals

- Data management plans (DMPs)
- Data sharing (in FAIR repositories)

Data protection: enforced measures

- European laws (e.g. GDPR)
- National laws (e.g. DPA/LDP)
- Cantonal laws
- Domain-specific laws (e.g. Federal Act on Research Involving Human Beings)

Personal and sensitive data

Personal data: all information relating to an identified or identifiable person (art. 3 lit. a FADP)

Very broad notion!

Sensitive data: Personal data on:

1. Religious, ideological, political or trade-union; related views or activities;
2. Health, the intimate sphere or racial origin;
3. Social security measures;
4. Administrative or criminal proceedings and sanctions

 (art. 3 lit. c FADP)

Data sharing in the social sciences

Better established for quantitative data

Stronger challenges/resistance when it comes to qualitative data:

- Complex nature of qualitative data
- Increased risks of identification
- Lack of know-know

Anonymisation is a central measure to protect research participants and allow data sharing

Defining qualitative data anonymisation

Qualitative data include:

Written information, but also voice or pictures

This presentation focuses on interview transcripts

What is qualitative data anonymisation?

Qualitative data anonymisation is like quantitative data anonymisation. It is about rendering research participants anonymous by removing identifying information from research data.

Data can be anonymised for at least two purposes:

- Publications
- Secondary use

Anonymisation – a definition

- The notion of anonymisation refers to a process by which the elements allowing the identification of a person are definitively deleted from a dataset, a document, an interview transcript, etc.
- Legally, this means that an individual cannot be identified without significant effort.
- Should key or raw data be kept, then the data are considered pseudonymised.
- Anonymisation represents a principal solution for complying with data protection requirements. Anonymised data that cannot be linked to a living individual is not subject to data protection acts.

Specificities of qualitative data anonymisation

As opposed to quantitative data anonymisation, qualitative data anonymisation:

- is often hard to achieve;
- involves somewhat different anonymisation techniques;
- should be done at specific times of the research life cycle;
- is more time consuming and costly;
- should be done by the research team

Anonymisation as a process

Three-layered approach

Sometimes anonymisation is impossible or affects the quality and re-use potential of the data. Consider anonymisation together with consent and access.

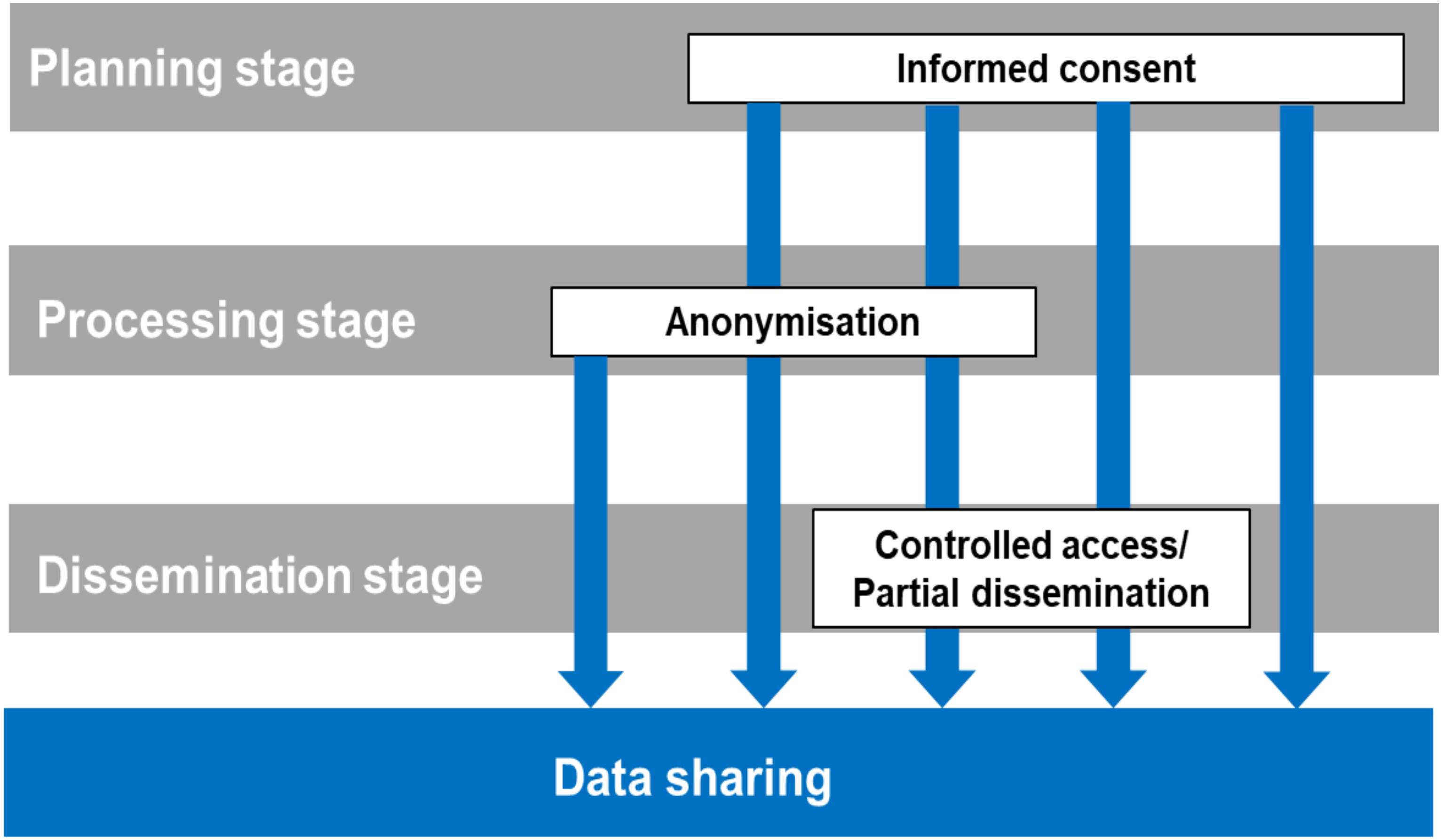
- Ask for consent to share personal data
- Regulate/restrict user access

Controlling access is a better option than over-anonymising

How can access be controlled?

Depending on the chosen repository, the following access conditions apply:

- Access without registration: data can be accessed by anyone
- Access upon registration: data can only be accessed by registered users
- Restricted access:
  - Access only for academic research and training
  - Access only for academic research
  - Access only granted upon request.
- Access after an embargo period


Planning stage
Informed consent
Processing stage
Anonymisation
Dissemination stage
Controlled access/
Partial dissemination
Data sharing

Some common forms of resistance

«Full anonymisation of qualitative data is impossible»

- Large volume of more or less subtle identifying information, which makes anonymisation very difficult.
- The removal of some identifying information can make identification difficult enough for the data to be legally considered as anonymised.
- Combined with other measures, anonymisation, even partial, can be a very effective layer of protection.

«I cannot share data that are not fully anonymised»

- Data that are not fully anonymised are considered personal data
- Sharing this type of data is subject to strict conditions in terms of information, consent, security and access.
- If only partial anonymisation is possible, it is important to combine this measure with others.

«Anonymisation makes the data worthless»

- Since anonymisation involves removing information, it naturally impoverishes the data.
- Depending on the type of re-use intended, anonymisation may be more or less problematic.
- Anonymisation implies a sacrifice, but it does not mean removing everything.
- A right balance must be found.
- Adding other layers of protection means more information can be kept.

«Anonymisation is not worth the effort»

- The fact that anonymisation is labour intensive and costly should not be a reason for not sharing the data.
- Make sure you budget resources at the time of the proposal.
- Make sure you set up an anonymisation strategy, including doing anonymisation at the right time during your research.

Proceeding to data anonymisation

Removing identifying information

The first step is to identify identifying information

- It is easier for quantitative data
- It can vary for qualitative data

Identity disclosure

A person’s identity can be disclosed through identifying information:

- A value may, possibly in combination with other values, lead to (re-)identification:
- A value is easily determined (e.g., by acquaintances)

Direct and indirect identifiers

Identifying information may consist of:

- direct identifiers, which alone are sufficient to identify people (e.g., name, AVS number);
- strong indirect identifiers, which allow fairly easy identification (e.g., home address, telephone number); and/or
- weak indirect identifiers, which allow identification through *combinations* of characteristics (e.g. socio-demographic and background information)

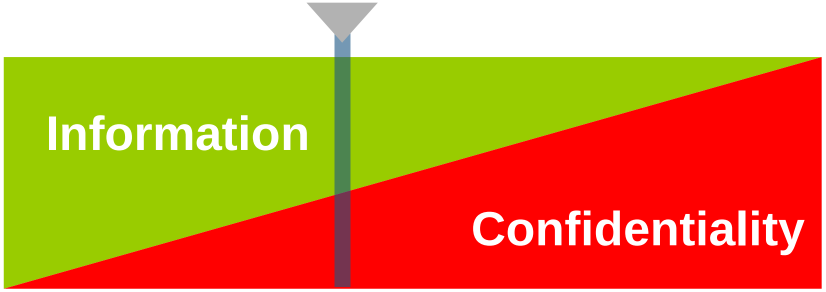
Anonymisation: main steps

As a general rule, the following principles apply:

- Direct identifiers are removed from the data;
- Strong and weak indirect identifiers are either removed or transformed;
- Documentation should also be verified.

Best conducted by the investigators/research team

How much anonymisation is enough:



Over-anonymising can distort data, make them unusable, unreliable or misleading

Anonymisation techniques

Qualitative anonymisation techniques include:

- replacing personal names with aliases
- categorising proper nouns
- changing or removing sensitive information
- categorising background information
- changing values of identifiers

Replacing personal names with aliases

Changing proper nouns into aliases is the most common anonymisation technique.

- It is always a better option to use aliases rather than simply delete the names or replace them by a letter [x];
- It is important to be consistent in the selection and use of aliases throughout a research project;
- The same aliases should be used in both the data and the published excerpts.

Categorising proper nouns

Names of people who have no essential importance in understanding the data content can be removed from the data without creating aliases. These names can be replaced with broader categories such as:

[woman], [man],
[sister], [father],
[colleague, female], [neighbour male]

The same may apply to other proper nouns, such as institutions, names of places, etc.

[Lower secondary school], [restaurant],
[hometown], [residential area]

Categorising proper nouns (2)

Large towns can usually remain [e.g., London].

Remember to check the background information as well (e.g., the name of a specific restaurant could help reveal the place of residence).

Changing or removing sensitive information

Identifying sensitive information should be removed, categorised or classified. For example, if relevant to the subject matter, a rare disease could be recoded to [severe long-term illness].

Removing sensitive data is justified if the respondent mentioned it incidentally, if the information is not relevant to the subject matter, or if it constitutes a disclosure risk.

Categorising background information

Background characteristics of participants such as gender, age, occupation, workplace, or place of residence are often essential for understanding the data.

Background data can be categorised or recoded using existing classifications.

Date of birth → [1991]
Elephant trainer → ???

Changing values of identifiers

Sometimes it is possible to anonymise qualitative data by distorting information, just like values of identifying attributes can be swapped between records in quantitative data.

Removing hidden metadata from files

Do not forget to check whether archival files contain any hidden technical metadata that could enable the identification of research participants. Technical metadata may be saved when files are created but also when they are edited.

Managing anonymisation

Plan anonymisation early in the research as part of your data management plan;

Develop an anonymisation strategy. This may include:

- File management rules
- Mandatory anonymisation (e.g. direct identifiers, places, ages and dates)
- Possible anonymisation (e.g. medical information, sensitive information)

File handling procedures

1. Experiment with anonymisation by processing a couple of files at first;
2. Make a copy of the unanonymised file and put it in a secure location;
3. Begin anonymising on the copy. Files should be named clearly so the anonymised version can be identified;
4. Use consistent procedures within a single file and throughout the project;
5. Use specific characters, such as [ ] for anonymisation to help keep track of what has been changed or not;

File handling procedures (2)

6. Document the anonymisation process. Keep an anonymisation log of all replacements, aggregations or removals made – keep separate from anonymised data files;
7. Utilise ‘find & replace’ to change names to their aliases;
8. When anonymisation is finished, erase original files and lists of aliases. Review the background material relating to the data because they may contain identifiers.

Tip: Plan or apply editing at the time of transcription. For example, mark each proper noun with a special character that is not used elsewhere (e.g. **#).**

Top 10 recommendations

Top 10 recommendations

1. Never promise anonymisation
2. Do not mix up anonymisation and pseudonymisation
3. Do not take anonymisation too lightly
4. Do not reduce anonymisation to direct identifiers
5. Do not over-anonymise
6. Ask for financial resources to conduct anonymisation
7. Plan for anonymisation
8. Develop an anonymisation strategy
9. Choose the right anonymisation techniques
10. Do not collect personal data if not needed

Resources

A few resources on anonymisation

- CESSDA Data Management Expert Guide (DMEG)
- Anonymisation and personal data, Finnish Social Science Data Archive: https://www.fsd.tuni.fi/en/services/data-management-guidelines/anonymisation-and-identifiers/
- FORS Guide: Data Anonymisation: Legal, Ethical, and Strategic Considerations
- Upcoming FORS Guides on data anonymisation techniques (quantitative and qualitative)

FORS GUIDES
to survey methods
and data management

Data anonymisation: legal, ethical, and strategic considerations

Alexandra Stam[1] and Brian Kleiner[1]

[1]FORS

FORS Guide No. 11, Version 1.0

June 2020

https://forscenter.ch/publications/fors-guides/

Want to learn more?

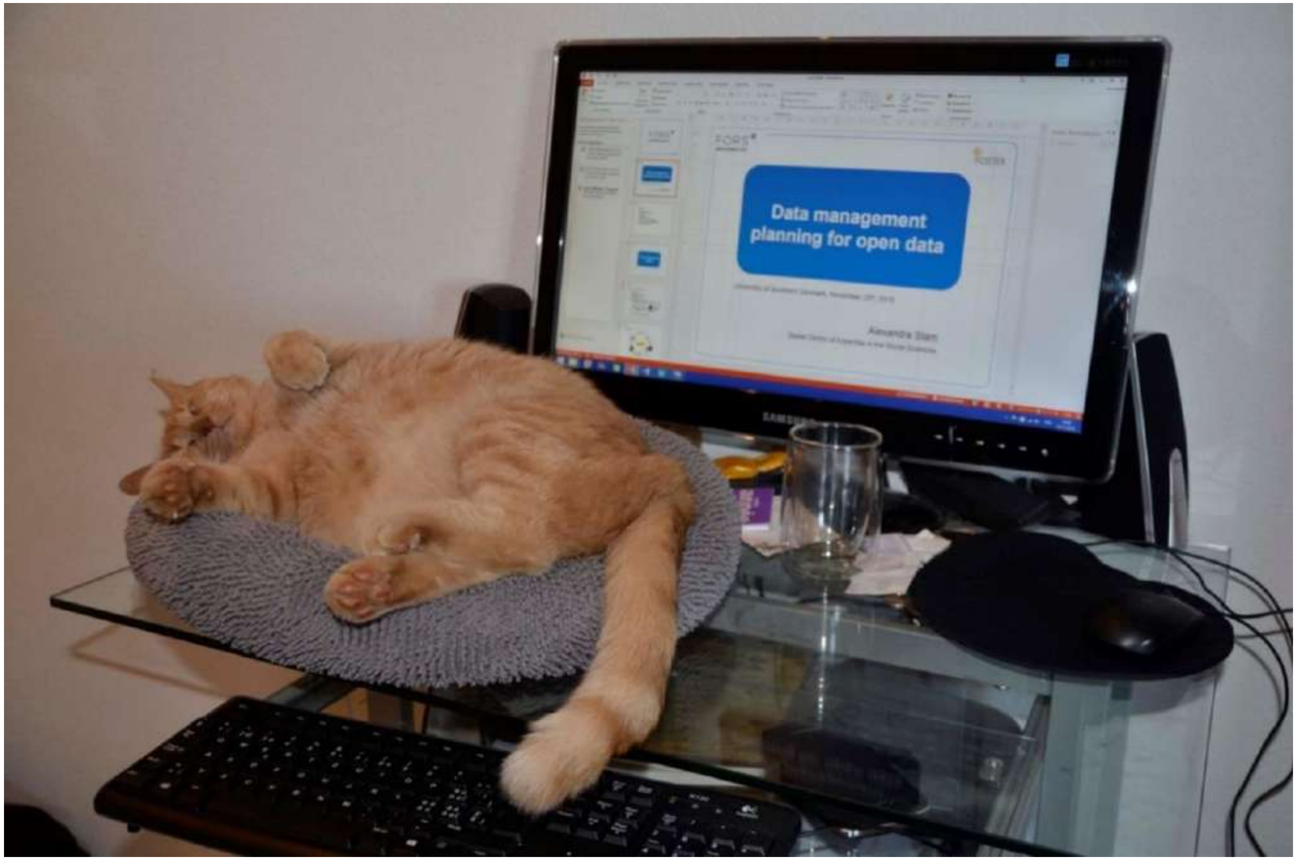
- Data archiving
- Data management support
- FORS guides
- SWISSUbase
- …

www.forscenter.ch/data-services/help-resources/

dataservice@fors.unil.ch

alexandra.stam@fors.unil.ch

Questions?


FORS
Data management planning for open data
Alexandra Stam
SAMSUNG