FORS data management webinar series

Quantitative data anonymisation

Dr. Brian Kleiner
November 1, 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 / November 22nd

FORS

Objective:

Transmit basic know-how for a sound anonymisation of quantitative research data

FORS

Outline

1. Background
2. Anonymisation strategy
3. Anonymisation techniques
4. Wrap-up

Background

FORS

New requirements

From funders:

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

From journals:

- Deposit of data used in publications
- Sufficient documentation

Anonymisation in the current research environment

- Digitalisation of data: more and more data are produced
- New research fields, including new types of data
- Facilitated access to data by the community
- New analytical/data extraction tools
- Computational power allows for analysis of increasingly rich datasets (incl. linked data)
- ‘Contradictory’ forces: protection and openness

Examples of data protection acts:

European level: the GDPR applies

National level: countries have their own legal bases (e.g., the UK Data Protection Act)

Swiss level: Federal Data Protection Act (DPA/LPD)

There are also cantonal laws

Note that other domain-specific laws may apply, e.g., Federal Act on Research involving Human Beings

FORS

Data protection acts: common elements

- The processing of personal and sensitive data is subject to legal basis
- There are exemptions for specified purposes and retention of personal data when processed for research
- Anonymised data that cannot be linked to a living individual is not subject to data protection acts

Personal data

Personal data means any information enabling direct or indirect identification of a human subject.

Most common in research are:

- Names and background information
- Information that can be traced online (social media)
- Information that is connected to a personal ID (administrative data)
- Audio and video

FORS

Sensitive data (special types of personal data)

Although definitions may change across cultures and legal frameworks, personal data are usually considered sensitive when they relate to the following topics:

- Racial or ethnic origin
- Political opinions
- Religious or philosophical beliefs
- Trade union membership
- Physical or mental health
- Criminal offences and court proceedings
- Intimate life
- Genetic data
- Biometric data

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.
- Anonymisation represents a principal solution for complying with data protection requirements.
- Legally, this means that an individual cannot be identified *without significant effort*.

Direct and indirect identifiers

- Direct identifiers alone are sufficient to identify people (e.g., name, AVS 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

Direct and indirect identifiers - example

Soc. Sec. Nr.	Gender	Age class	Region	Education	Profession	Income
1927384123	Female	40-55	Zurich	Higher	Civil Servant	80'000
1927384124	Male	30-40	Pully	Middle	Fisherman	50'000
1927384125	Male	55+	Vers-chez-les-Blanc	Higher	Politician	250'000
1927384126	Male	20-30	Yverdon	Lower	Plumber	70'000
1927384127	Female	55+	Lutry	Higher	Surgeon	150'000
1927384128	Male	30-40	Aubonne	Higher	IT consultant	80'000
1927384129	Male	55+	Zurich	Unknown	Surgeon	160'000
1927384130	Female	20-30	Corcelles	Middle	Violin Maker	60'000
1927384131	Female	30-40	Neuchatel	Lower	House cleaner	55'000
...	...	...	...	...	...	...
...	...	...	...	...	...	...

FORS

Anonymisation strategy

FORS

Factors to be considered in your strategy

1) The nature and type of personal data to anonymise
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

One other factor - resources

The extent to which you anonymise, and the techniques that you choose will be influenced by the **resources** you have available to carry out the work.

If you apply for a grant from the SNSF, you can budget up to 10,000 CHF for data management, including for anonymisation of data.

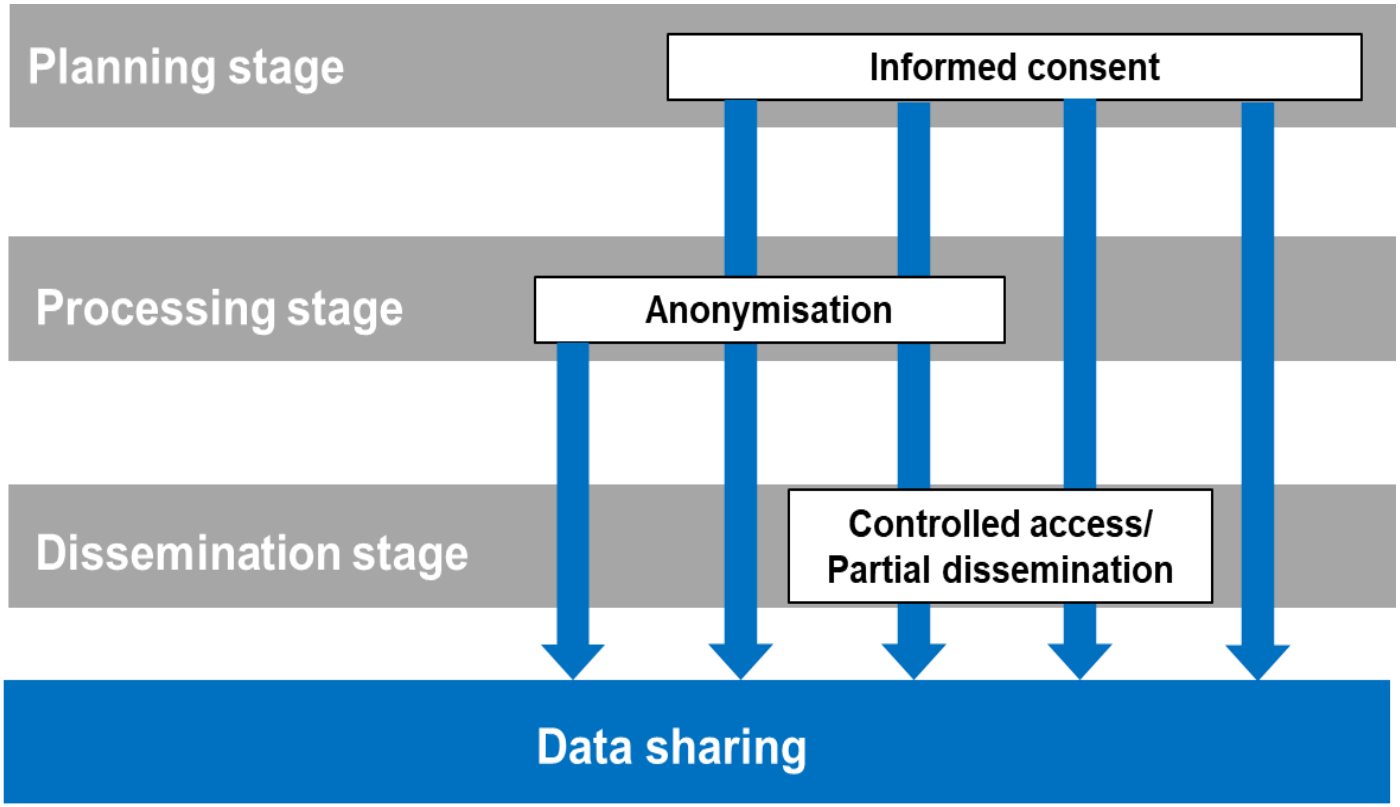
Developing your anonymisation strategy

Your strategy should be developed early in the project and include at least:

- an evaluation of disclosure risk, and
- a description of the anonymisation measures and their rationale.

The strategy will serve as documentation for secondary users. Its implementation should be described after anonymisation has been completed.

Three-layered approach



FORS

Anonymisation techniques

FORS

About anonymisation techniques

Techniques are ways of removing, masking, or modifying data in order to make it more difficult to identify individuals in a file.

The techniques you choose to apply should be driven by your overall anonymisation strategy.

Choosing techniques

To select the appropriate techniques, ask the following questions with respect to your strategy:

- What types of direct or indirect identifiers do my materials contain? Is there rare/unique information in the data?
- What combinations of variables or information can allow identification of an individual?
- What characteristics of the data do I want to retain (if possible) and which ones can be “sacrificed” in the anonymisation process?

Based on the answers to these questions (as well as the risks identified beforehand), you will be able to decide which data to delete, edit, categorize, and so on.

Some general principles and considerations

- Different techniques are appropriate with different types of variables
- Different techniques modify the dataset and its variables and records in different ways
- Risk should be reduced to an acceptable level
- Give preference to lighter techniques, if possible
- Choosing the appropriate technique requires expertise with the subject matter
- Each technique has advantages and disadvantages

Basic approach

- Removal of direct and particular strong indirect identifiers;
- assessment of weak indirect identifiers and appropriate techniques;
- starting with a categorisation table.

Categorisation of variables

Identifier type	Direct identifier	Strong indirect identifier	Indirect identifier
Social security number	x		
Full name	x		
Email address	x	x	
Phone number		x	
Postal code			x
District/part of town			x
Municipality of residence			x
Region			x
Major region			x
Municipality type (urban, semi-urban, rural)			x

FORS

Key specific anonymisation techniques

- Variable suppression
- Record suppression
- Character masking
- Pseudonymisation
- Generalisation
- 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 and direct identifiers

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., generalisation)
- In some cases you can just suppress or alter a value for a variable within a record (e.g., an outlier)

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

Pseudonymisation

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

Pseudonymisation - example

Before anonymisation:

Person	Pre Assessment Result	Hours of Lessons Taken Before Passing
Joe Phang	A	20
Zack Lim	B	26
Eu Cheng San	C	30
Linnie Mok	D	29
Jeslyn Tan	B	32
Chan Siew Lee	A	25

After pseudonymising the Person attribute:

Person	Pre Assessment Result	Hours of Lessons Taken Before Passing
416765	A	20
562396	B	26
964825	C	30
873892	D	29
239976	B	32
943145	A	25

FORS

Generalisation

- Reduction of precision of a variable
- Create discrete categories from a continuous variable (e.g., age, income)
- Recoding
- Combine string values into broader categories (e.g., profession)

Generalisation – example

Commune	District
Aclens	Morges
Agiez	Jura - Nord vaudois
Arnex-sur-Orbe	Jura - Nord vaudois
Arzier-Le Muids	Nyon
Assens	Gros-de-Vaud
Ballaigues	Jura - Nord vaudois
Belmont-sur-Lausanne	Lavaux-Oron
Belmont-sur-Yverdon	Jura - Nord vaudois
Cheseaux-Noréaz	Jura - Nord vaudois
Jongny	Riviera - Pays-d'Enhaut
Jorat-Mézières	Lavaux-Oron
Moiry	Morges
Penthalaz	Gros-de-Vaud

Data perturbation

- Modification of values to be slightly different
- Where small changes of values do not significantly affect analysis and accuracy in most cases
- An example is base-x rounding

Example – base-x rounding

Person	Height (cm)	Weight (kg)	Age (years)	Smokes?	Disease A?	Disease B?
198740	160	50	30	No	No	No
287402	177	70	36	No	No	Yes
398747	158	46	20	Yes	Yes	No
498732	173	75	22	No	No	No
598772	169	82	44	Yes	Yes	Yes

Person	Height (cm)	Weight (kg)	Age (years)	Smokes?	Disease A?	Disease B?
198740	160	51	30	No	No	No
287402	175	69	36	No	No	Yes
398747	160	45	18	Yes	Yes	No
498732	175	75	21	No	No	No
598772	170	81	42	Yes	Yes	Yes

FORS

Wrapping up

FORS

Putting it all together – applying the strategy

1. Determine the access conditions for your data
2. Determine the acceptable re-identification risk threshold and expected utility
3. Categorise variables in the dataset
4. Remove undesirable variables
5. Anonymise direct and indirect identifiers – apply techniques
6. Assess actual risk and compare against threshold
7. Perform more anonymisation, if necessary
8. Evaluate the solution again
9. Document the anonymisation process

A few tips for best practice

1. Minimisation – ask only for what you need in collection
2. If possible, use syntax in statistical software
3. Follow good practice in storage and data security
4. Be consistent across waves (if longitudinal)

Conclusion

- Always consider anonymisation of research data together with consent agreements and access restrictions;
- Regulating/restricting user access may offer better solution than fully anonymising;
- Remove, mask, change direct and indirect identifiers;
- Maintain maximum information to the extent possible;
- Plan anonymising at start of research, not at the end

Want to learn more?

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

www.forscenter.ch

dataservice@fors.unil.ch

brian.kleiner@fors.unil.ch

FORS

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

FORS

A few resources

- CESSDA Data Management Expert Guide (DMEG)
- Guide to Basic Data Anonymisation Techniques (https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Other-Guides/Guide-to-Anonymisation_v1-(250118).pdf)
- FORS Guide: Data Anonymisation: Legal, Ethical, and Strategic Considerations
- Upcoming FORS Guides on data anonymisation techniques (quantitative and qualitative)

FORS webinar series

1. Informed consent / September 27th
2. Data documentation / October 11th
3. Quantitative data anonymisation / November 1st
4. Qualitative data anonymisation / November 22nd

FORS

Questions or comments?

FORS