Data anonymisation

Brian Kleiner – FORS

Outline

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

Background

Anonymisation in the current research environment

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

Anonymisation in data management

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

Anonymisation – a definition

- Anonymisation refers to the process by which the elements in data are definitively deleted or modified, making identification impossible, or as least extremely difficult, thus complying with data protection requirements.
- Fully anonymised data are no longer personal data.

Pseudonymisation

- Refers to the removal or replacement of identifiers with pseudonyms or codes, which are kept separately and protected by technical and organisational measures.
- The data remain pseudonymous as long as the original identifying information exists.
- Pseudonymised data remain personal data.

https://www.fsd.uta.fi/aineistonhallinta/en/anonymization-and-identifiers.html

Anonymisation – a difficult promise

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

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
- …

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
…	…	…	…	…	…	…
…	…	…	…	…	…	…

Anonymisation strategy

Anonymisation: *part* of the solution

Applying anonymisation techniques to data is key to addressing the tension between ORD and data protection.

However, advising researchers to rely entirely on data anonymisation may be ill-advised:

- anonymisation strips value from data
- (full) anonymisation is often not feasible in practice
- researchers may lack the know-how to do it properly
- researchers may not have the resources to do it properly

The need to focus on a larger data protection/sharing strategy

Rather than rely exclusively on anonymisation, a more holistic approach to data sharing should take into account:

- the nature of the data and potential harm to participants
- the future use of the data and purposes – potential analytic utility
- informed consent
- sharing goals, and future access conditions
- resources

Developing an anonymisation strategy

An anonymisation 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.

General principles and considerations

- Different anonymisation techniques are appropriate with different types of data.
- Different anonymisation techniques modify the data in different ways.
- Risk should be reduced to an acceptable level.
- Preference to lighter techniques.
- Choosing the appropriate techniques requires expertise with the subject matter.
- Each technique has advantages and limitations.

Quantitative data anonymisation

About quantitative 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 selected techniques should be driven by the overall anonymisation strategy.

Choosing techniques

To select the appropriate techniques, researchers should ask the following questions with respect to their strategy:

- What types of direct or indirect identifiers do their 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 they 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), they will be able to decide which data to delete, edit, categorize, and so on.

Key specific quantitative anonymisation techniques

- Variable suppression
- Record suppression
- Character masking
- Pseudonymisation
- Generalisation
- Data perturbation

Qualitative data anonymisation

What is qualitative data anonymisation?

Anonymisation of qualitative data tends to be more complex than anonymisation of quantitative data.

Data are anonymised for at least two purposes:

- In publications when citing what was said
- Sharing data for secondary analyses or teaching

Quantitative anonymisation techniques

Anonymisation techniques for qualitative data include:

- Replacing personal names with aliases
- Categorising proper nouns
- Changing or removing sensitive information
- Categorising background information
- Changing values of identifiers

Wrapping up

Conclusion: some advice to researchers

Don't assume that data cannot be shared – there are creative solutions that are compatible with data protection legislation.

Plan anonymisation as part of sharing strategy at the start of a research project, not at the end.

Always consider anonymisation of research data together with consent agreements and future access restrictions.

Regulating/restricting user access may offer a better solution than fully anonymising.

Maintain maximum information to the extent possible.

A few resources for anonymisation

- 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 GUIDES

to survey methods
and data management

- Ethics in the era of open research data: some points of reference (FORS Guide N°03), Pablo Diaz
- The informed consent as legal and ethical basis of research data production (FORS Guide N°05), Sybil Krügel
- How to draft a DMP from the perspective of the social sciences, using the SNSF template (FORS Guide N°07), Pablo Diaz, Alexandra Stam
- Pre-registration and registered reports (FORS Guide N°09), Marieke Heers
- Data anonymisation: legal, ethical, and strategic considerations (FORS Guide N°11), Alexandra Stam, Brian Kleiner
- Replication in the social sciences (FORS Guide N°16), Marieke Heers
- Data protection: legal considerations for research in Switzerland (FORS Guide N°17), Pablo Diaz
- Data Citation: How and Why Citing (Your Own) Data (FORS Guide N°19), Christina Bornatici, Nicolas Fedrigo
- Qualitative data anonymisation: theoretical and practical considerations for anonymising interview transcripts (FORS Guide N°20), Alexandra Stam, Pablo Diaz
- Data Sharing in the Social Sciences (FORS Guide N°21), Marieke Heers

FORS GUIDES
to survey methods
and data management

- Ethics in the era of open research data: some points of reference (FORS Guide N°03), Pablo Diaz
- The informed consent as legal and ethical basis of research data production (FORS Guide N°05), Sybil Krügel
- How to draft a DMP from the perspective of the social sciences, using the SNSF template (FORS Guide N°07), Pablo Diaz, Alexandra Stam
- Pre-registration and registered reports (FORS Guide N°09), Marieke Heers
- **Data anonymisation: legal, ethical, and strategic considerations (FORS Guide N°11), Alexandra Stam, Brian Kleiner**
- Replication in the social sciences (FORS Guide N°16), Marieke Heers
- Data protection: legal considerations for research in Switzerland (FORS Guide N°17), Pablo Diaz
- Data Citation: How and Why Citing (Your Own) Data (FORS Guide N°19), Christina Bornatici, Nicolas Fedrigo
- **Qualitative data anonymisation: theoretical and practical considerations for anonymising interview transcripts** (FORS Guide N°20), Alexandra Stam, Pablo Diaz
- Data Sharing in the Social Sciences (FORS Guide N°21), Marieke Heers

Questions?