FORS GUIDES
to survey methods
and data management

Quantitative data anonymisation: practical guidance for anonymising social science data

Brian Kleiner[1] and Marieke Heers[1]

[1] FORS

FORS Guide No. 23, Version 1.0

March 2024

Abstract:

In the social sciences, requirements from funders and journals to make data available often present difficulties for researchers because of data protection issues. Anonymisation is a good solution for addressing the challenges of personal and sensitive data. This FORS Guide provides some practical guidance on how to select and apply techniques for anonymising quantitative data within a larger strategic framework for sharing.

Keywords: data sharing, data protection, data management

How to cite:

Kleiner, B. & Heers, M. (2024). Quantitative data anonymisation: practical guidance for anonymising sensitive social science data. *FORS Guide*, 23, Version 1.0, 1-17. https://doi:10.24449/FG-2024-00023

The 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.

Editorial Board:

Emilie Morgan de Paula (emilie.morgandepaula@fors.unil.ch)
Florence Lebert (florence.lebert@fors.unil.ch)

FORS, Géopolis, CH-1015 Lausanne
www.forscenter.ch/publications/fors-guides
Contact: info@forscenter.ch

Acknowledgement:

This Guide is based largely on discussions, presentations, and trainings conducted by the Data Management Services (DMS) team at FORS. We would especially like to thank Alexandra Stam and Pablo Diaz for their guidance, and Emilie Morgan de Paula for her many thoughtful suggestions on the content as well as the form. Moreover, we thank Céline Racine for a very helpful review of the Guide.

Copyright: 

Creative Commons: Attribution CC BY 4.0. The content under the Creative Commons license may be used by third parties under the following conditions defined by the authors: You may share, copy, freely use and distribute the material in any form, provided that the authorship is mentioned.

1. INTRODUCTION

With the strengthening of the open science movement, researchers are under increasing pressure from funders and academic journals to make the data underlying their analyses available to the larger scientific community, either for re-use, replication, or teaching purposes. However, in the social sciences these demands often present difficulties because of data protection requirements, as well as ethical considerations. Most data involving human subjects cannot easily be shared without certain treatment or processing (Late & Kekäläinen, 2020). The application of anonymisation techniques plays a key role in this context as it allows for making available data that otherwise could not be shared. Thus, anonymisation is growing in importance, among other data management skills that pertain to data sharing.

Researchers often find themselves in between the seemingly contradictory demands for data sharing on the one hand, and data protection requirements on the other. Yet, there is little available to guide researchers in how to approach this complex issue in practice, and specifically on how to apply anonymisation techniques appropriately to quantitative data in the social sciences. This FORS Guide provides an overview of the key considerations in selecting and applying anonymisation techniques to quantitative data, within a larger strategic framework regarding data protection and data sharing. The first section sets the stage by presenting relevant definitions, followed by a treatment of the larger strategic considerations around data protection and the proper role of anonymisation in the context of data sharing. The main part focuses on the selection and application of key anonymisation techniques to quantitative data in the social sciences. This is followed by some conclusions and practical recommendations. The present guide aims at providing guidance to social science researchers with respect to anonymising their social science quantitative data – yet we emphasise that each project is unique, and that the researcher (or research team) is best qualified to make the right decisions for his or her specific project.

This FORS Guide is the third in a series on data anonymisation. The first Guide (Stam & Kleiner, 2020) focuses on the larger legal, ethical, and strategic considerations around anonymisation, while the second (Stam & Diaz, 2023) addresses the anonymisation of qualitative data, in particular interview transcriptions. For a general introduction to the benefits and requirements of data sharing, see the FORS Guide 21 (Heers, 2023).

2. ABOUT QUANTITATIVE DATA ANONYMISATION

2.1 DEFINITIONS

Before considering anonymisation strategy and techniques, some key definitions are in order. Most are presented in the other FORS Guides on anonymisation (namely, FORS Guide 11 (Stam & Kleiner, 2020) and FORS Guide 20 (Stam & Diaz, 2023)), but are repeated here for ease of reading.

Personal data and sensitive data

Article 5 of the Swiss Federal Act of Data Protection of 25 September 2020 (Status as of 1 September 2023) (FADP; RS 235.1) defines personal data as “any information relating to an identified or identifiable natural person”. The FADP further defines sensitive personal data (art. 5(c)) as:

"1. data relating to religious, philosophical, political or trade union-related views or activities,

2. data relating to health, the private sphere or affiliation to a race or ethnicity,

3. genetic data,

4. biometric data that uniquely identifies a natural person,

5. data relating to administrative and criminal proceedings or sanctions,

6. data relating to social assistance measures;"

The processing of sensitive data is prohibited under the FADP, with various exceptions, including scientific research. Anonymised data are no longer personal or sensitive data, and therefore the FADP does not apply.

Direct and indirect identifiers

Direct identifiers include information that is sufficient on its own to identify an individual. Examples include a person's full name, an e-mail address containing a person's full name, Old-Age and Survivors' Insurance (OASI) number (i.e., French/Italian AVS, German AHV), biometric identifiers, a person's voice, or a picture.

Strong indirect identifiers allow the identification of someone with high probability, either on their own or in combination with other information (in the dataset or elsewhere). Examples include postal address, phone number, vehicle registration number, web address to a page containing personal data, a unique professional position, an unusual job title, a very rare disease.

Weak indirect identifiers do not allow identification of someone on their own, but they could lead to identification when linked with other information. Examples include socio-demographic and background variables such as age, date of birth, gender, education level, occupation, income, marital status, mother tongue, place of work, area of residence, etc. Researchers working with social science data should pay particular attention to weak indirect identifiers that in combination risk identification of a respondent.

Anonymisation

First, we define anonymisation as a process by which the elements allowing the identification of a person are definitively removed from data and related documentation, such that an individual cannot be identified without significant effort. This definition corresponds to the legal definition of anonymisation, as stipulated in most data protection laws, including the FADP, in the sense that it is permanent and irreversible and involves a strong protection threshold: identification should no longer be possible, or only with very intensive effort.

Often, deidentification is treated as synonymous with anonymisation, although the sense of these terms may vary depending on relevant applicable data protection laws and regulations. Anonymisation is more consequential, as the data cannot be linked to an individual, while deidentification implies that explicit identifiers are removed (for a detailed explanation see Chevrier et al., 2019).

Pseudonymisation

Anonymisation should be distinguished from pseudonymisation, which consists in the removal or replacement of identifiers with pseudonyms or codes, where the identifiers are retained

separately and secured by technical and organisational measures. Data remain pseudonymous as long as the original identifying information is somehow kept by the researcher or his or her institution, which legally is the owner of the data. This often applies to longitudinal data, where contact lists must be kept for future data collection waves. Unlike anonymised data, pseudonymised data remain "personal" for the holder of the related identifying keys and are therefore subject to legal obligations. On the other hand, if researchers share pseudonymised data without the related identifying keys, then those data are considered anonymous for the recipients.

2.2 ANONYMISATION AS PART OF A PROJECT'S DATA PROTECTION STRATEGY

It is important to consider the larger context around data protection and the strategies to be employed in its favour. For more detailed information about the legal, ethical, and strategic aspects of data anonymisation, see our FORS Guide dedicated to this (Stam & Kleiner, 2020). For the purposes of the current Guide, what is important to keep in mind is that it is the legal and ethical responsibility of the researcher to ensure that steps are taken to make identification of study participants at least highly difficult. This can mean the application of specific anonymisation techniques, but also other measures that help to minimise the risk of identification and harm, such as restricting access to the data. The steps taken also depend on the consent – what has been promised to respondents must be respected.

In general, in formulating a strategy for protecting study participants, researchers should take into account both legal and ethical concerns, including an assessment of the risks of harm for study participants from potential disclosure, as well as a consideration of the effects of anonymisation on the potential utility of the data. When developing an anonymisation strategy, researchers need to identify for a specific project the right balance between these two.

Anonymisation is only one tool among several available to researchers for protecting respondents, and it should be developed and applied within the framework of the project's overall Data Management Plan (DMP). A good data protection strategy will select and combine the most appropriate measures given the nature of the study and the data that are to be collected. Besides anonymisation, this includes informed consent and data access controls.

In developing an effective strategy for data protection for a project, researchers should ask themselves these questions (for more details, see FORS Guide 11 (Stam & Kleiner, 2020)):

- What should be promised to respondents regarding the future use of their data?
- What is the nature and type of the personal data to anonymise? How difficult will it be to adequately anonymise the data?
- How sensitive are the data? What harm might be caused to respondents if they are identified?
- Who will be the future users of the data? Will usage be limited to researchers? What are the chances of improper use?
- What will be the likely uses of the data in the future? What level of data utility will be required in order to address these uses?

The answers to these questions will inform your strategy and the combination of protection measures put into practice for a project. These questions also illustrate that data protection, including anonymisation, needs to be reflected upon from the beginning of the project.

Your strategy for data protection should then incorporate three elements that will allow you to find the appropriate balance between data openness, utility, and protection: 1) informed consent, 2) level of anonymisation, and 3) access controls. In general, the greater the risk of harm to respondents, the more of each that should be applied, that is, more anonymisation, stronger promises regarding data protection and clarity about future uses of the data, and stricter controls on who can access the data and under which conditions. The *three-layered approach* developed by FORS shows how these three elements (consent, anonymisation, and controlled access) allow for the sharing of most social science research data (for a detailed description see FORS Guide 21 (Heers, 2023)).

3. ANONYMISATION TECHNIQUES

3.1 ABOUT THE TECHNIQUES

Anonymisation techniques are ways of removing, masking, or modifying data in order to make it impossible or extremely difficult to identify individuals in a dataset. This refers to the data file itself, but also to the accompanying documentation. The techniques you choose to apply should be driven by your overall anonymisation strategy and, as described above, will also depend on the consent obtained from respondents, as well as the access conditions that will be put in place.

To select the appropriate techniques, researchers should ask themselves the following general questions:

- What direct or indirect identifiers do the research materials contain? Is there rare/unique information in the data?
- What combinations of variables or information could allow identification of an individual?
- What characteristics of the data should be retained (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, researchers can decide which data to delete, edit, categorise, and so on. Below are several general principles and considerations regarding the selection of techniques:

- Different techniques are appropriate with different types of variables.
- Different techniques modify the data file and its variables and records in different ways.
- With the selected application of techniques to the direct and indirect variables it is important to mitigate the risk of identification or unauthorised disclosure of personal information to the point where the remaining risk is considered acceptable from a legal and ethical perspective. This means finding the right balance between privacy protection and data utility.
- Give preference to lighter techniques. When choosing anonymisation techniques, it is important to prioritise where possible less aggressive ones, as long as privacy is sufficiently protected. This means those that have a lower impact on data utility and potential secondary analyses, preserving as much as possible the original data's structure and granularity.

- Choosing the appropriate techniques requires expertise with the subject matter. Domain knowledge is key in understanding the specific variables and their interrelations. An effective anonymisation approach should rely on a comprehensive understanding of the data, their context, and associated privacy risks.
- There is no on-size-fits-all solution – each technique has its own benefits and drawbacks.

3.2 PRELIMINARY STEPS

Categorisation of variables

An important first step in the process of anonymising quantitative data is the categorisation of the variables in a data file as direct or indirect identifiers. This will lay the foundation for the subsequent stages of the anonymisation process, enabling the creation of a privacy-preserving and analytically useful dataset. This includes the following:

1. Data inventory: start by creating an inventory of all variables in the dataset. This will provide a comprehensive view to ensure that no variables are overlooked.
2. List direct identifiers: review each variable to identify those that contain direct personally identifying information (e.g., full names, OASI number), and classify them as such.
3. List strong indirect identifiers: review the remaining variables and identify those that could identify an individual with high probability, e.g. a postal address or a unique professional position.
4. Identify weak indirect identifiers: for variables that are neither direct nor strong indirect identifiers, assess whether these might be combined with other variables to significantly increase the chances of identifying individuals in the file. Based on this, consider which of the indirect variables could be modified to different extents, keeping in mind future possible uses of the data by secondary users. Here it is important to not only look at the variables but also to inspect their values.
5. Where there is any doubt, consider consulting with data privacy or domain experts with respect to the risks of identification with different combinations of indirect variables.
6. Document your categorisations in a table, as illustrated in Table 1. You can annotate the variables with particular notes for you or your project team to keep in mind. The table will serve as a sort of dashboard, where you can document further the techniques that you will apply to each of the direct and indirect variables. We encourage you to share this table when depositing your data, so that secondary users of your data can comprehend the data processing steps.

Selection of anonymisation techniques for each targeted variable

Given your overall anonymisation strategy and the general principles and considerations presented above, make decisions for each variable concerning how it should be treated by applying the techniques described in the following sections.

Table 1. Categorisation table for variables in a dataset.

Identifier type	Direct identifier	Strong indirect identifier	Weak indirect identifier	No identifier	Notes
Full name	X				Remove
OASI number	X				Remove
Email address	X				Remove
Phone number		X			Remove
Postal code			X		Remove
Municipality			X		Remove
Canton			X		Keep
Profession			X		Categorise
Education level			X		Categorise
Age			X		Categorise
Favourite colour				X	Keep as is

Note. Designed by the authors for illustration.

3.3 KEY TECHNIQUES

Variable suppression

Variable suppression involves the removal of entire variables from a dataset, where these contain personally identifiable information. It is employed most often in the case of direct identifiers and strong indirect identifiers that could easily lead to individual respondents (e.g., names, email addresses, telephone numbers, and open-ended text responses). Since this technique is the most drastic, in that it represents the most extreme loss of information in the data, it should be done as a last resort. Also, variable suppression should be the first applied to the data when applying the techniques.

An example where variable suppression is an effective method comes from the Swiss Household Panel (SHP) (SHP Group, 2023): Information on respondents' commune is removed from the distributed dataset, as only few analyses require this level of detail. Secondary users of the data who need this variable for their analyses can submit a special request which is assessed by the research team. If evaluated positively, they receive the data with information on the commune.

- Advantages: Variable suppression fully protects respondents by not revealing the personal information contained in direct identifiers or certain strong indirect identifiers.
- Disadvantages: Suppressing entire variables can reduce the utility and richness of the original data. When you collect longitudinal data, you cannot permanently delete all identifying information because respondents need to be contacted again in each subsequent data collection wave.

Record suppression

Record suppression means removing from a data file an entire individual record that cannot easily be anonymised. Some respondents report extreme values across one or more variables that make it easier to identify them without disproportionate effort. For example, one can imagine a case where a respondent has reported having 12 children, earning 10 million francs per year, and working as a high-level official in a cantonal administration. Such a combination would argue for a complete erasure of that respondent from the file.

Naturally, removing an individual entirely is extreme in itself and reduces to some extent the analytic utility of the data. In such cases, one must assess whether other techniques where values are modified rather than erased could adequately reduce risks of identification. Considering the previous example, the researcher could also decide to recode the values: the same respondent would then fall into the following categories: > 4 children; salary of >250'000 francs/year; and working in administration. In that case, generalisation is applied, a technique explained in more detail below.

In some cases, rather than suppressing a record, you can just remove (i.e., recode into missing) or alter a specific value for a variable within a record (e.g., an outlier). An example for removing a value comes from the 10th wave of the Swiss European Social Survey (ESS) (ESS; European Social Survey, 2023): When a respondent lives in a household where more than 10 children live, the birth year and his or her relationship with the respondent are set to "No answer/Not available". An example for altering a specific value comes from the same data: the variable indicating respondents' age is recoded in a way that values higher than 90 are altered to 90. Here, generalisation, and more specifically top coding, is again applied.

Before applying record suppression, make sure that lighter techniques (such as generalisation) cannot be applied.

- Advantages: Record suppression fully protects individuals that stand out too much in a file and is effective particularly in high-risk scenarios.
- Disadvantages: It can reduce the richness of the data, especially concerning abnormal values, as well as the representativeness of a sample.

Generalisation

The technique of generalisation involves reducing the precision of a variable, replacing specific data points with broader, less precise categories or ranges, thus making it more difficult to identify individuals, especially when combining different indirect variables. This can mean creating a set of categories to replace the individual values of a continuous variable, or collapsing existing response categories of a nominal or ordinal variable, this often when there are too few respondents in a given "cell" within a category.

To take an example for the former, with the continuous variable "INCOME", where exact figures are obtained from respondents, these can be attributed to categories, as in:

1.	35,560	→	30,000-39,999
2.	21,120	→	20,000-29,999
3.	52,999	→	50,000-59,999
4.	120,600	→	More than 100,000
5.	79,005	→	70,000-79,999

Categories should be created such that the responses are roughly evenly distributed, so that there are no categories with too few cases. The latter would increase chances of identification if combined with other variables and would restrict the options for data analysis in some cases.

String values of a continuous variable can also be generalised, for example, where more specific professions are placed into higher-level categories:

1.	Software engineer	→	Information technology
2.	Nurse	→	Healthcare
3.	Teacher	→	Education
4.	Security analyst	→	Information technology
5.	Accountant	→	Business

This is similar to the above example, where the “high-level official in a cantonal administration” is recoded into working in “administration”. The above-mentioned case where the number of 12 children is recoded into “> 4” is another example of generalisation.

In the social sciences, it is common to analyse occupations. To do so in a structured way, the International Standard Classification of Occupations (ISCO) has been established (International Labor Organization, 2023). If you work with ISCO-codes and your variable consists of four digits, i.e., provides information on unit groups (e.g., Generalist Medical Practitioners), you can generalise the variable by recoding it into a 3- (Medical doctors), 2- (Health Professionals), or 1-digit-code (Professionals). Whenever possible, it is advisable to categorise data according to established and tested categories, ontologies, or vocabularies to increase semantic interoperability.

Another example is the removal of the level of detail from a date of birth. Instead of displaying the full date of birth one might only keep the birth year.

As an example of generalising of a nominal variable, consider the following: Imagine that a survey asks respondents about the postal code of their residence, yet as we have seen above, this variable, in combination with others (e.g., profession, country of origin), can be a weak indirect identifier. Generalisation from postal code to a typology of communes (Federal Statistical Office, 2023) allows reduction of the risk of identification, while keeping a high degree of granularity. Another example is recoding countries of origin into larger geographical regions.

As a general rule, the application of the generalisation technique should be applied to indirectly identifying variables in such a way that it minimises chances of identification of individuals while maintaining to a certain extent analytic potential. Again, it is up to the researcher(s) to find the right balance.

- Advantages: Generalisation allows for retention of survey variables and analytic utility.
- Disadvantages: It reduces the precision of original data, limiting to some extent what can be done for analyses. For some statistical analyses categorical variables are more demanding than continuous ones. Since the data are modified but not eliminated, generalisation does not guarantee that respondents cannot be identified when variables are crossed.

Top coding

Top coding is a specific form of generalisation, as you have seen in some examples above. It is useful to hide outliers, such as extremely high salaries or old age, without having to remove a full record. For example, in the Swiss edition of the European Social Survey (ESS) round 10 data (Ernst Stähli et al., 2023), people aged older than 90 have been recoded as 90 years old. Households with more than 20 members have been recoded as "20 members" (European Social Survey, 2023). Decisions for such top coding must consider the analytic implications and should be documented.

Generalising open text questions in surveys

In social science surveys it is common to collect some information in an open format, that is, respondents are not presented with any pre-defined categories but can express themselves by providing a text response. This is often the case for professions. In the context of anonymisation, it is important to pay special attention to these variables as they often include identifying information. A commonly applied anonymisation strategy is the categorisation of open text questions. For example, in youth surveys young adults are sometimes asked about their parents' profession in an open format. If they respond that the mother works at the UBS bank, this can be categorised as "banking".

- Advantages: This allows collection of accurate information.
- Disadvantages: With categorising open texts, there is an extreme loss of information. If the information needs to be reduced, it might be more efficient to collect data in a closed way right away. In any case, researchers should reflect upon the implications on anonymisation before collecting the data.

Character masking

Character masking is when a fixed or variable number of characters are covered over or "masked" by symbols that render the response unidentifiable, but where some of the original characters are maintained. This process ensures that individual identities remain hidden, while the masked data still retain some of their properties for research purposes. The specific pattern of masking can vary, such as masking a fixed number of characters or only displaying the first few characters.

To take an example, the following (made up) OAIS numbers could be masked by randomly replacing 6 of the 13 digits with an "X":

1.	7890123456789	→	78X01X345X89X
2.	2345678901234	→	2X4567X90X23X
3.	4567890123456	→	45X78X0X45X6
4.	8901234567890	→	8X0X23X678X0
5.	1234567890123	→	1X34X67X90X3X

It is important to note that the level of character masking can be adjusted depending on the sensitivity of the data and the specific requirements for anonymisation.

- Advantages: Character masking preserves the format and length of the original data, making it less disruptive to data analysis and processing. This ensures that any

algorithms or systems relying on the data structure can still function properly. Masked data can still be useful in performing data validation and quality checks without revealing sensitive details. This is essential to ensure the accuracy and consistency of the data. In some cases, masking data can provide better data usability than completely redacting the information or suppressing variables. Researchers can still recognise what the original data looked like, and secondary data users can understand what information is being masked.

- Disadvantages: Improper application of character masking might lead to identification, when combined with other variables.

Pseudonymisation

With pseudonymisation, identifying information is replaced with artificial identifiers or codes, removing direct associations with individual respondents. For instance, personally identifiable data such as names or email addresses can be replaced with unique, randomly generated strings or numerical codes. In the following example, full names are replaced by unique and randomly generated identification numbers:

1.	Lukas Müller	→	9012
2.	Anna Zimmermann	→	4823
3.	Jean-Paul Lévesque	→	6821
4.	Laura Schneider	→	8402
5.	Matteo Bianchi	→	3207

Statistical programmes, such as Stata, R, and SPSS, can easily produce such IDs.

As noted earlier, pseudonymisation is usually employed when it is necessary to keep original identifying information about respondents, for instance for longitudinal studies where one must return to respondents at a later date. As such, pseudonymisation must always be reversible. To do this, a key that links the original and pseudonymised data should be kept in a secure and separate location. Most statistical analysis software programs (e.g., R, SPSS, Stata) allow for pseudonymisation of data, as well as special dedicated anonymisation software tools.

Pseudonymisation can be applied to different levels of the data. For example, think of pupils in the same class, attending the same school. While usually researchers are not interested in the name of the class or the teacher, they might want to examine if there is a class or teacher effect on certain outcomes. In such cases, it makes sense to replace the name of the teacher with a unique identifier. Another example is individuals living in the same household in a specific neighbourhood. In the Swiss Household Panel, for example, each household has an id (idhous*), where individuals living in that household can clearly be matched to that household via their individual id (idpers*). This allows the data analyst to account for the nested structure of the data. The list with respondents' contact details is kept separately from the data by another institution.

- Advantages: Pseudonymisation enables recontact of respondents at a later date, as well as data linkage across datasets. It also allows grouping individuals without revealing the group's identifying information.

- Disadvantages: Respondents may still be vulnerable to identification, since the original personal identifying information is kept. With pseudonymised data, data protection laws still apply.

Data perturbation

Some techniques introduce random noise or alterations to original data values, without significantly compromising the overall statistical properties, integrity, or accuracy of the data. One such technique is data perturbation, where small random adjustments are made systematically to numerical values. This ensures that the modified data closely resemble the original data with respect to distribution, while making it more challenging to identify individuals in the data. Data perturbation is also called randomization.

An example is base-x rounding, which adds random noise by rounding values to a nearest multiple of x. In the following, real age values undergo base-3 rounding:

1.	23	→	24
2.	46	→	45
3.	19	→	18
4.	72	→	72
5.	58	→	57

Another common type of data perturbation is data swapping, where specific values of a variable are swapped between different records within the data file (for an example see Table 2). Data swapping is also referred to as permutation. Usually this is done with care to maintain important relations in the data, especially with respect to certain demographic characteristics, such as level of education or age. In the following example, income is swapped only between records that share the same level of education:

Table 2. Respondent income and education before and after swapping.

Respondent	Income	Education
Before swapping		
A	23,000	Bachelor
B	45,000	Master
C	29,000	High School
D	57,000	Master
E	36,000	Bachelor
After swapping		
A	36,000	Bachelor
B	*57,000*	Master
C	29,000	High School
D	*45,000*	Master
E	23,000	Bachelor

Note. Underlined numbers and those in italics refer to swapped data.

As another example, consider that a researcher has collected data on the compositions of households. In very few households, there are twins; here, data perturbation could be applied and the birth year of one of the twins could be changed by a year.

In Table 3 and based on Kasprzak et al. (2023), we present another example of data swapping: In the table, respondents' birth dates have been altered by ±15 days. An alternation of 15 days might be sufficient for some purposes but not others -- this depends on the combination of variables and other factors such as the population studied. When applying data swapping researchers have to rely on their expertise.

Table 3. Altered birthdates by ±15 days, as an example for data swapping.

Original dataset	Swapped dataset
Birthdate	*Birthdate*
17.02.1980	02.02.1980
07.04.1974	22.04.1974
25.10.1993	10.10.1993
06.07.1978	21.07.1978

Note. Designed by the authors based on Kasprzak et al. (2023).

- Advantages: Data perturbation disrupts any direct link between respondents and their potentially identifying data, while maintaining analytic utility and the statistical properties of the data.
- Disadvantages: Excessive perturbation may distort variables and compromise the accuracy of analyses, while insufficient perturbation may lead to the risk of identification, especially if variables are combined.

Be aware that even as a data producer you cannot always anticipate all potential reuse possibilities of your data. While you might assume that a variable can be swapped without impacting the results of the analysis, this might not hold for other analyses. It is particularly important to document the data perturbation you apply.

3.4 APPLYING THE TECHNIQUES

By way of summary, given the overall strategy and particular selected techniques for a data file, follow these steps in practice:

Step 0: Think about anonymisation from the beginning of the project. For each step, consider how the decisions taken will impact the of the data.

Step 1: Categorise the variables in your data file as direct identifiers, strong indirect identifiers, or weak indirect identifiers, and create an inventory table of variables requiring treatment.

Step 2: Remove or modify direct identifiers (e.g., suppression, masking, pseudonymisation).

Step 3: For high-risk records check if a technique lighter than removal can be applied. Otherwise, remove the record.

Step 4: Apply remaining techniques to targeted indirect identifiers (e.g., generalisation, perturbation), with care for the proper order, given possible interactions.

Step 5: Assess actual risk of identification in the resulting data file against acceptable risk threshold, in relation to access conditions and potential breach and harm. Iteratively improve the anonymisation until the risk falls within the acceptable threshold.

Step 6: Assess overall utility of the data and whether more precision could be reintroduced into the file. Confirm that the data still meet the intended use and analysis requirements.

Step 7: Document all decisions and the entire process. Integrate documentation into materials released with the anonymised data.

4. CONCLUSION

Anonymisation is a crucial tool when preparing your data for sharing. Depending on the nature of the data, different techniques can be applied. Anonymisation must be considered throughout the research project, and an anonymisation strategy must be developed taking other factors into account, such as the risks for participants, as well as informed consent and access conditions. There is not one "right" anonymisation strategy that fits all research projects - each project is unique and needs to identify the right balance between protecting respondents and data utility. This must be done against the background of what has been promised to respondents (consent) and the way the data will be made available for secondary analysis. Expertise on the topic is required to define an adequate anonymisation strategy. Secondary users of the data might want to use the data for purposes that have not been considered at all by the data producer. Therefore, researchers should be careful with techniques that alter the data and should document their procedures well.

5. RECOMMENDATIONS

Here we provide key recommendations for things to keep in mind when anonymising quantitative data:

Recommendation 1 – Keep in mind that anonymisation is difficult to achieve with social science data.

Recommendation 2 – Plan anonymisation at the beginning of your research project, and not at the end. This will allow you to avoid pitfalls that would slow down or prevent its appropriate application.

Recommendation 3 – Always consider anonymisation of research data together with consent agreements and access restrictions, with respect to potential risk and data utility. If anonymisation has been promised to respondents, this promise must be kept. We recommend not promising anonymisation, as it is difficult to achieve with social science data. Better use a formulation such as "you will not be identifiable in the data".

Recommendation 4 – Regulating/restricting user access may in some cases offer a better solution than full anonymisation, where data utility may be too diminished.

Recommendation 5 – In order to avoid unnecessary operations to protect respondents, one good practice is to ask only for what is really needed in one's data collection (i.e., minimisation). Collecting data that afterwards must be suppressed or manipulated might involve an

unnecessary burden for respondents. Therefore, with respect to anonymisation, we encourage you to consider the consequences of your data collection instrument while you design it.

Recommendation 6 – Try to maintain maximum information in the data to the extent that this is possible.

Recommendation 7 – Use syntaxes in statistical software to apply the anonymisation techniques. This not only saves time but also helps you to document the anonymisation process.

Recommendation 8 – During data collection and data processing, follow good practice in data storage and security, ensuring that only eligible people can access the data.

Recommendation 9 – If you are doing longitudinal research, be sure to be consistent in how anonymisation is done across waves.

Recommendation 10 – Make available your data only in trusted digital data repositories. We recommend SWISSUbase. This refers to anonymised as well as non-anonymised data.

6. FURTHER READINGS AND USEFUL WEB LINKS

Chapter 5 from the CESSDA Data Management Expert Guide on "Anonymisation" developed by the CESSDA Training Team (2017-2022) provides a good overview and interesting examples on quantitative data anonymisation.

The FORS Guide by Stam and Kleiner (2020) on "Data anonymization: legal, ethical, and strategic considerations" is a good basis to learn more about the theoretical underpinnings and challenges related to data anonymisation in the social sciences.

The FORS Guide by Stam and Diaz, (2023) on "Qualitative data anonymisation: theoretical and practical considerations for anonymising interview transcripts" is a valuable resource for anonymising qualitative social science data.

Kasprazak et al.'s (2023) article is an excellent reference on data swapping.

REFERENCES

CESSDA Training Team. (2017-2022). *CESSDA Data Management Expert Guide – Online tool developed by the Consortium of European Social Science Data Archives (CESSDA) to help researchers make their data FAIR*. CESSDA ERIC. https://dmeg.cessda.eu/

Chevrier, R., Foufi, V., Gaudet-Blavignac, C., Robert, A., & Lovis, C. (2019, May 31). Use and Understanding of Anonymization and De-Identification in the Biomedical Literature: Scoping Review. *Journal of Medical Internet Research, 21*(5), e13484. https://doi.org/10.2196/13484

Ernst Stähli, M., Ochsner, M., Pollien, A., Nisple, K., & Sapin, M. (2023). *European Social Survey, Switzerland - 2021 (Round 10) (Version 1.0.0) [Data set]* FORS data service. https://doi.org/doi.org/10.48573/8p06-0t58

European Social Survey. (2023). ESS round 10 - 2020. Democracy, Digital social contacts. https://ess-search.nsd.no/en/study/172ac431-2a06-41df-9dab-c1fd8f3877e7

Federal Act on Data Protection of 25 September 2020 (Status as of 1 September 2023) (FADP; RS 235.1), (2020). https://www.fedlex.admin.ch/eli/cc/2022/491/en

Federal Statistical Office. (2023). *Typologie de communes en 22 classes au 5.12.2000 (Recensement)*. https://www.bfs.admin.ch/bfs/en/home/statistics/catalogues-databases/maps.assetdetail.461374.html

Heers, M. (2023). Data Sharing in the Social Sciences. *FORS Guides, 21*. https://doi.org/10.24449/FG-2023-00021

International Labor Organization. (2023). *International Standard Classification of Occupations (ISCO)*. ILO. https://ilostat.ilo.org/resources/concepts-and-definitions/classification-occupation/

Kasprzak, J., Frey, S., Oetlinger, H., Westphalen, C. B., Erickson, N., Heinemann, V., & Nasseh, D. (2023, 2023/01/01). Swapping data: A pragmatic approach for enabling academic-industrial partnerships. *DIGITAL HEALTH, 9*, 20552076231172120. https://doi.org/10.1177/20552076231172120

Late, E., & Kekäläinen, J. (2020). Use and users of a social science research data archive. *PloS one, 15*(8), e0233455. https://doi.org/10.1371/journal.pone.0233455

SHP Group. (2023). *SHP 2023. Swiss Household Panel. Living in Switzerland Waves 1-23 + Covid 19 data (Ref. 932 ; Version 6.0.0) [Data set]* FORS data service. https://doi.org/10.48573/642z-p311

Stam, A., & Diaz, P. (2023). Qualitative data anonymisation: theoretical and practical considerations for anonymising interview transcripts. *FORS Guides, 20*. https://doi.org/10.24449/FG-2023-00020

Stam, A., & Kleiner, B. (2020). Data anonymization: legal, ethical, and strategic considerations. *FORS Guides, 11*. https://doi.org/10.24449/FG-2020-00011