



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

Alexandra Stam¹  and Pablo Diaz² 

¹ FORS

² FORS-UNIL

Abstract:

This guide addresses qualitative data anonymisation from a theoretical, legal and practical perspective. After defining qualitative data anonymisation and addressing some common forms of resistance when it comes to anonymising such data, this guide presents the main anonymisation techniques. It argues for a layered approach and encourages the reader to consider anonymisation as an important layer of protection amongst others.

Keywords: anonymisation techniques, qualitative data, interview transcripts, data protection, identity disclosure, identifiers.

How to cite:

Stam, A, Diaz, P. (2023). *Qualitative data anonymisation: theoretical and practical considerations for anonymising interview transcripts*. FORS Guide No. 20, Version 1.0. Lausanne: Swiss Centre of Expertise in the Social Sciences FORS. doi:10.24449/FG-2023-00020

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)

Giannina Vaccaro (Giannina.Vaccaro@unil.ch)

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

Acknowledgement:

We would like to thank Arja Kuula-Luumi (Finnish Social Science Data Archive – FSD) for her most helpful comments and feedback on an earlier version of this Guide. Thanks also to Brian Kleiner and Emilie Morgan de Paula (FORS Data Service) for their precious inputs throughout the writing of this 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

Within the context of Open Science, researchers are increasingly required to share their research data. While the Open Data principles are generally well accepted by the scientific community, the sharing of the empirical materials may raise some practical challenges. This is particularly true for researchers working with qualitative data.

One reason for this is that qualitative data pose particular challenges in terms of participant protection. The richness and complexity of this type of data and the conditions under which they are produced often make it difficult to implement privacy protection measures, including anonymisation.

This guide, which complements a previous FORS Guide on the legal, ethical and strategic considerations around anonymisation ([Stam & Kleiner, 2020](#)), addresses qualitative data anonymisation, with a special focus on interview transcripts. The first part defines qualitative data, and presents anonymisation as a technique for protecting research participants that is even more effective when combined with other measures. The second part addresses some common forms of resistance when it comes to the anonymisation of qualitative data and how to overcome them. The third part discusses anonymisation from a practical point of view, and presents the most common techniques used for qualitative data. Finally, this Guide concludes with some recommendations on how to best plan for and implement qualitative data anonymisation.

2. ABOUT QUALITATIVE DATA ANONYMISATION

2.1 DEFINING QUALITATIVE DATA

The anonymisation of qualitative data is a complex issue. The very notion of qualitative data is problematic in the sense that there is no such thing as data that would be, by nature, strictly qualitative. Data are always (co)constructed and are, therefore, the result of a process of selection, formatting and framing. To speak of qualitative data is therefore to speak of data produced, processed and/or interpreted within the framework of a so-called “qualitative” approach, as opposed to a “quantitative” one. In other words, qualitative data designate materials used to validate the results of a research project in a non-numerical logic (DuBois, Strait, & Walsh, 2018). While all data can be processed qualitatively or quantitatively - as shown by the growing success of lexicometric methods in discourse analysis (Rioufreyt, 2019) - there are materials that are most often associated with qualitative approaches. This is notably the case with:

- interviews and focus groups transcripts;
- verbatim notes;
- fieldnotes;
- life history calendars;
- documents: reports, meeting minutes, e-mails;
- video recordings;
- audio recordings;

- pictures;
- diaries.

This Guide focuses on one of the most common types of data used in qualitative research, namely interview transcripts.

2.2 DEFINING QUALITATIVE DATA ANONYMISATION

The Federal Act on Research involving Human Beings defines anonymized data as data “which cannot (without disproportionate effort) be traced to a specific person” ([art. 3, let i HRA](#)). In other words, for the data to be legally considered anonymous, it must be impossible, or at least extremely difficult, to trace them to specific individuals. This means permanently removing all identifying information from empirical materials. If the process is reversible, e.g. if another non-anonymised version of the data is kept, the data should not be considered as anonymous but still as personal. Thus, if the data are only pseudonymised – i.e. the identifying information is replaced by codes and stored separately in a secure manner – their processing remains subject to data protection laws.¹ The consequences of this are outlined in the next section.

Data can either be selectively anonymised for publication purposes, for example by altering interview excerpts in such a way as to make them non-identifying, or extensively anonymised for archiving and sharing purposes, for example by anonymising a whole dataset. While anonymising data extracts for publications is a common practice, anonymising all empirical materials for secondary use is less common and faces a number of barriers, some of which can be addressed by a better understanding of the topic and some practical solutions.

The general principles of anonymisation are the same for quantitative and qualitative data (see [Stam & Kleiner, 2020](#)). Put simply, it is about removing direct identifiers and assessing indirect identifiers, some of which will need to be modified or else removed (see section 4.2 for a definition of direct and indirect identifiers). There are, however, some important differences between the anonymisation of quantitative and qualitative data which are worth mentioning here. To list a few, in comparison with quantitative data, qualitative data anonymisation:

- **is often complex:** it may be a labour-intensive process due to the heterogeneity and often the significant amount of data collected. Furthermore, it may even be difficult sometimes to identify all potentially identifying information;
- **involves somewhat different anonymisation techniques;**
- **should be done at specific points of the research life cycle:** anonymisation, or at least the identification of the information to be anonymised, should ideally be carried out at the time of transcription, and not at the end of the project (as is often the case with quantitative data anonymisation);
- **is much more time consuming and costly:** hence the importance of proper planning at the earliest stage (ideally in the data management plan) in order to be able to request resources and to identify when anonymisation should take place in the research life cycle;

¹ In most cases, from the perspective of someone who does not have access to the code and for whom it would be impossible to identify an individual on the basis of the information in his or her possession, the data can be considered anonymous. This is not true, however, for data that are subject to the Human Research Act (Erard, Heusghem & Parisato, 2021).

- **requires deep understanding of the data and the associated risks:** anonymisation is about making choices on what to keep or not, and this can best be done by the research team, who best know the data, the associated risks for research participants, and the value of the data for reuse purposes.

2.3 ANONYMISATION AS A PROCESS: TOWARDS A LAYERED APPROACH

Anonymisation is a powerful tool, and if applied correctly allows for bypassing of data protection laws, since anonymised data no longer contain personal information. However, full anonymisation, or at least a very high level of anonymisation, is very hard to achieve. This is particularly true for qualitative data, which often focus on lived experiences in very situated contexts. When working with materials such as interview transcripts, it is therefore preferable to assume that personal data are being processed. Therefore, it is necessary to comply with data protection laws, in particular by providing participants with adequate and complete information that takes into account the disclosure of their data to third parties. We invite you to consult our FORS Guide "Data protection: legal considerations for research in Switzerland" ([Diaz, 2022](#)) for more information on this topic.

Based on these considerations, we consider anonymisation primarily as a process that tends towards the definitive removal of all potentially identifying information, but whose result may be (more or less) partial. Our understanding of anonymisation in this Guide is therefore more nuanced and more practically-oriented than the stricter legal definition. More precisely, we take anonymisation to be a key layer of protection (deployable to different degrees) that should be combined with other measures such as informed consent or data access control to protect the individuals involved in research projects. For example, if research participants have given their informed consent for their data to be shared after having made it clear that their full anonymity cannot be guaranteed, but that the main information leading to their possible identification will be removed, then it is possible to apply a lower level of anonymisation. If, on the other hand, there may be significant risks for research participants should their identity be revealed, then it is possible to add an extra layer of protection by controlling who will have access to the data. Some repositories, such as [SWISSUbase](#), allow for access control. For example, this could involve limiting the use of the data to research purposes and only by those affiliated with a research institution. For data that present greater risks, it could involve granting access on a case-by-case basis only upon prior approval of the data producer. It is therefore always preferable to combine different measures in order to retain as much information as possible in the end. This will be further illustrated in the following section.

3. ADDRESSING SOME COMMON FORMS OF RESISTANCE TO ANONYMISATION

This section addresses a few common forms of resistance to anonymisation that need to be nuanced. Understanding some key obstacles and ways to encounter them allows one to better envision when and how to carry out qualitative data anonymisation.

“Full anonymisation of qualitative data is impossible”

It is true that data used in research based on qualitative methodologies often contain a large volume of more or less subtle identifying information. Full anonymisation then becomes almost

impossible, since participants may remain recognizable even when deleting all directly identifying variables.

However, the difficulty of achieving total anonymisation of data collected and/or used in qualitative research does not make the process irrelevant. Indeed, the removal of some identifying information can make identification if not impossible at least difficult enough to be highly discouraging. When coupled with other measures (consent, access control, etc.), even partial anonymisation can be a very effective layer of protection. For example, sometimes only specific people (relatives, friends, etc.) can recognize an individual on the basis of certain information. If it can be ensured that these people do not have access to the data, the chances of re-identification are greatly reduced - even if the data legally remain personal data.

“I cannot share data that are not fully anonymised”

Data that are not fully anonymised must, for legal purposes, be considered personal data. Nonetheless, sharing this type of data is still possible, with appropriate conditions in terms of information, consent, security, and access. For more information, see our FORS Guide on data protection in Switzerland ([Diaz, 2022](#)).

When it comes to archiving and/or sharing personal data, there is no standard solution or procedure. In cases where only partial anonymisation is possible, it is important to combine this measure with others, such as obtaining appropriate consent and/or controlling access to data (user agreements, limitation of rights, embargos, etc.). Note, however, that it is never appropriate to share personal data that may put participants at high risk of harm.

“Anonymisation makes the data worthless”

Since anonymisation involves removing information, it naturally impoverishes the data. For this reason, it is often perceived to result in too great a loss of value of empirical materials. That said, depending on the type of reuse intended, as well as the epistemological posture of the "secondary" researchers, anonymisation may be acceptable or problematic to different degrees.

The reuse of research data may have significantly different purposes from those intended by the original project. Some of these purposes may be less sensitive to the removal of identifying information. This may be the case for:

- re-analysis for teaching purposes (methodology) and focus groups transcripts
- re-analysis for history of science purposes
- research design analysis (inspiration for future research)

Depending on the epistemological posture of the researcher, the anonymisation of materials may be more or less problematic. For example, approaches close to anthropology are highly context-centred, which makes the removal of potentially identifying elements undesirable (Swiss Anthropological Association (SAA), 2021, p. 7). The same is true for historical approaches for which sources must be traceable in order to be scientifically valid (Müller, 2006). While these arguments are absolutely relevant and should be considered, it should be emphasized that reusing data usually does not involve addressing the same questions with the same analytic approaches as the researchers who produced them. Data produced within a specific epistemological and methodological framework can be reused with other tools, questions and perspectives. Some approaches are, for example, less sensitive to contextual elements and can accommodate anonymous data (Tenevic & Weber, 1992).

Finally, anonymisation implies a sacrifice of detail, but it does not mean removing everything. A proper balance must be found. Even if the data cannot be used with the same level of detail as in the original analyses, they may still have interesting potential.

"The risks of re-identification are too great"

Information technology is constantly evolving. With the improvement in computing capacity and the growth of databases, the ability to recognize people is increasing day by day. What may be considered anonymous today may not be tomorrow. This is why it is important to thoroughly assess the level of anonymisation needed so that an individual cannot be recognized without a significant effort, as well as to consider adding other layers of protection such as informed consent and access control to lower the risks of re-identification. If research participants agreed for their data to be shared under certain conditions, then less anonymisation may be needed. If, on the other hand, no proper consent was obtained, then a stronger level of anonymisation needs to be applied. There may be cases where participants agreed for their data to be shared, but where identity disclosure may bear significant consequences for the research participants due to the sensitivity of the data, and where strong anonymisation would distort the data too much. For such cases, the Swiss National Science Foundation (SNSF) accepts that there may be ethical reasons for which data cannot be shared, and these reasons need to be described in the Data Management Plan (DMP).

"Qualitative anonymisation is too labour-intensive"

It is true that the anonymisation of qualitative data is labour-intensive, but it should not be a reason on its own for not anonymising and sharing the data. Indeed, there are ways to mitigate the needed efforts. Most notably, additional resources for anonymisation can be requested from the SNSF at the time of the research grant proposal, which offers up to 10'000 Swiss francs for data preparation. Also, good planning and implementation of anonymisation can help to save time. For example, one can plan for anonymisation by assigning and distributing responsibilities for it to project staff, and by defining the best moment for proceeding to anonymisation during the project. In addition, one can prepare anonymisation guidelines to ensure that anonymisation is done in a systematic and standardized way, especially if several team members carry out the anonymisation of the research materials.

"Anonymising people is not ethical"

Some researchers - especially those working with indigenous communities - object to anonymisation on the grounds that it tends to make certain populations invisible, while researchers publish with their names (Svalastog & Eriksson, 2010). While this remark, which reflects a genuine concern for social justice, is relevant, it should be nuanced by an assessment of risks. Indeed, it is not uncommon for people who do not wish to be anonymized - for example, political activists - to insist that their names be published. In these cases, it is important to evaluate the risks linked to the disclosure of participants' identities. Anonymisation should not be seen as something automatic. It must be thought out and weighed.

To sum up, while the above common forms of resistance are well-founded, they can often be overcome by considering anonymisation together with informed consent and access control. This allows for the retention of more information (and therefore value to the data) while meeting legal requirements when it comes to the sharing of personal data.

4. PROCEEDING TO DATA ANONYMISATION

As noted before, anonymisation is about removing or transforming personal data, in such a way that a person can no longer be recognized, or at least not without significant effort. Ideally, already at the time of collection, best practice is to avoid collecting personal data that are not essential to the research project. However, when the collection of personal data is necessary for the research project - as is often the case with some qualitative methodologies - or unintentional, then anonymisation techniques need to be applied. In this section, we will first present how to recognise direct and indirect identifiers, since these (potentially) identifying sorts of information are at the very heart of the anonymisation process. We will then present the concept of data minimisation, and how anonymisation efforts can be lowered when personal data collection is anticipated. Finally, we will then look into qualitative anonymisation techniques, as well as file handling procedures.

4.1 ASSESSING DIRECT AND INDIRECT IDENTIFIERS

A first step for anonymising qualitative data is to assess the type of information that will be collected, or has been collected, and to classify potentially identifying information into three main groups:

- **Direct identifiers:** 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, AVS number, biometric identifiers, a person's voice, or a picture, etc.
- **Strong indirect identifiers:** 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 position, unusual job title, a very rare disease, etc.
- **Weak indirect identifiers:** weak indirect identifiers allow identification in a non-obvious way when combined with other information. Weak indirect identifiers do not allow identification of someone on their own, but could lead to identification when linked with other information. Examples include socio-demographic and background variables such as age, date of birth, gender, education, occupational status, income, marital status, mother tongue, place of work, area of residence, etc.

One should not forget that research data may also contain identifiers relating to other individuals, such as friends, family, an employer, and so on. This information must also be anonymised.

While the assessment of direct and indirect identifiers often occurs after the data have been collected, it is strongly encouraged to start reflecting on the data that may be collected before the interviews start, as a way to avoid unnecessarily collecting identifying information.

4.2. DATA MINIMISATION AS A FIRST ACTION TO FACILITATE ANONYMISATION

Even when the interview grids are very structured, it may happen that personal data that are not strictly essential for the analysis are collected. This may result from researchers' field strategies, which may involve an informal discussion at the beginning of the interview to create a climate of trust, or it could be completely out of one's control, with for example particularly chatty research participants wanting to share personal information that goes beyond the

bounds of the research project. Important to keep in mind is that the more personal data are being collected, the more complex and time-consuming the anonymisation work will be. Therefore, it is important to limit the collection of such non-essential data. This is known as the principle of data minimisation.

Data minimisation involves being aware of what type of information needs to be collected, and conversely what should not be collected in order to reduce the risk of obtaining unwanted information. One way is to explicitly ask research participants during the consent procedure not to provide personal information about themselves or others during the interview, such as full names, dates, and institution names. It is also possible to only start the recording after a first round of introduction has taken place, when personal information is often shared. If the research project requires the collection of socio-demographic information (e.g. age, salary, type of employment), it is useful to save it for the end of the interview, stop the recording, and use a pre-printed structured form with closed categories that follow existing social classifications. Indeed, avoiding open-ended questions allows for better control of the collection. In cases where research participants provide information that goes beyond the scope of the project, one should, if appropriate, redirect the discussion to focus on the research questions. Also, as noted by the [Finnish Social Science Data Archive \(FSD\)](#), the interviewer should avoid asking questions that will likely produce responses requiring heavy anonymisation measures.

4.3 DATA ANONYMISATION TECHNIQUES

Once data have been collected, and possibly disclosing information has been identified and assessed (ideally already at the time of data transcription), then anonymisation can take place.

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 (see below).
- Documentation related to the research methodology and data collection should also be verified, as it may contain identifying information. Should the research participants provide materials as part of the research methodology, it is important to check that the latter do not provide any potentially identifying information, including hidden metadata.

While the techniques in themselves may be rather straightforward to apply, the main difficulty consists in choosing what to keep and what to remove or transform. In the end, it is often an issue of balance, perspective and individual sensibilities about what is important to be retained, and what could be permanently removed or modified. The process may result in negotiations within the team. Should some potentially identifying element be considered as important to keep, then it may result in having to give up on other potentially interesting ones. The sometimes difficult process of decision-making has been well-described from the perspective of a qualitative data archive (Bendjaballah et al., 2018) and that of a research team (Saunders, Kitinger, and Kitinger, 2015). The final decision must take into account the nature of the harm if someone were to be recognized, other measures in place to comply with data protection policies, as well as the reuse potential of the data. Important to note is the fact that there is often significant diversity between the different interviews of the same research corpus. This means that single interviews may include more or less personal and sensitive information. As a result, to meet the anonymisation threshold it may be that an entire section, comprising too much personal information, needs to be removed from the transcription. In some other cases,

the research team may decide to remove an entire interview, which is too difficult to anonymise, from the corpus. It is always better to share part of the collected data than none.

The [Finnish Social Science Data Archive \(FSD\)](#) has identified five main techniques when it comes to the anonymisation of qualitative data in interview transcripts:

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

Each of these techniques is described below.

Replacing personal names with aliases

Replacing personal names with aliases is the most common anonymisation technique. Proper nouns can refer to the research participants, but also to other people they may refer to during the interview. It is always better to use an alias rather than simply delete names or replace them with a special character such as [x], as it makes the transcript easier to read and closer to the original version. Also, it makes it easier to track the same individuals cross a text.

While the original data may contain both a first and last name, the standard procedure consists in replacing the first names with aliases and removing the surnames. Since first names may provide contextual information, such as social, ethnic or religious origins (Coulmont, 2011), a standard procedure consists in choosing aliases which convey similar background information. For example, if the research participant is called John, we may rather give him the alias Mark, rather than Mohammed. In France, a tool has been developed that makes it possible to find socially related first names, based on the results obtained in the baccalaureate exams.² If a person is referred to by his or her surname only, and provided this is relevant for the research project and understanding of the data, then the alias should also be a surname. Should a dataset include the name of a person who is publicly known, then the name can be kept, provided that references to this person are made with respect to his or her public role, and not private life, in which case an alias must be used.

Should the data be part of a longitudinal study, it is important to stay consistent and use the same alias across waves. Also, the same aliases should be used in both the data and the published excerpts.

Categorising proper nouns

Names of people (or sometimes even the names of animals) who do not play an important role in understanding the data can be categorized rather than replaced with aliases. For example, depending on the interview and the information needed for best understanding the data, we may decide to change a woman's name to categories such as [a woman], [the respondent's sister], [the respondent's colleague], [a mother], a [female neighbour], etc.

The same may apply to other proper nouns, such as names of institutions (restaurants, schools, companies), places, street names, and so on. So, following the same logic, a restaurant's name could be replaced with more generic information, such as [an Italian

² <https://coulmont.com/bac/>

restaurant], [a local restaurant], or [a family-owned restaurant]. For example, the restaurant 'fleur de sel' could be described as a 'semi gourmet' restaurant, or 'a local restaurant' depending on its importance in the research context, as its name provides information about the locality, a small Swiss town. The name of a restaurant could indeed reveal the place of residence and increase the risk of identity disclosure.

The choice to categorise or not depends on the research context and the prevalence of other potentially identifying information. For example, one might choose not to categorise big cities but only small towns, if the risks of identifying someone are very low.

As a rule, the names of larger cities, such as Zurich or Geneva, can usually remain, but the names of smaller towns should be categorized. This will again depend on the research context, and the prevalence of other potentially identifying information.

Changing or removing sensitive information

Sensitive data are a particular category of data that need to be handled with care. If you collected them incidentally and they are not crucial information for the project, then they should be removed or at least categorized. If, however, these data are essential, then you could keep them, but in order to compensate you would need to remove other potentially identifying information. For example, if a person belongs to a civil disobedience movement, such as for example extinction rebellion in the Swiss context, and committed illegal acts, then you may decide to categorise it as 'militant activities'. This is particularly true when a person has carried out an illegal activity in an authoritarian country. Another example could be someone suffering from a rare disease. In this case, rather than stating the name of the disease, one could change it to [severe rare disease] or [long-term illness].

Categorizing background information

When studying people's behaviours, socio-demographic characteristics and background information, such as gender, age, occupation, and dates (e.g., of birth, death, marriage, graduation, arrivals or departures) are often important to make sense of the data and allow data reuse. It is possible to categorize such information, for example, by using national or international classifications, such as major regions at the Swiss level, or an international standard classification of occupations. What matters is to document which classification has been used, whether it is an established one, or one that was created as part of the project. For information such as age, it may be preferable to use categories, such as [20-25] and [more than 60], again depending on the importance of the information for making sense of the data. Categorising data is often a better solution than deleting such information.

Changing value identifiers

There may be situations when it is better to distort information by altering values rather than removing the information. For example, if you interviewed someone who participated in high-level sports competition, and if that is of importance, then you may change other values. For example, you could change the years when the competition took place, the region (place, country) or even the gender of the person, if this is not crucial information.

4.4 FILE HANDLING PROCEDURES

Doing qualitative data anonymisation is not only about applying techniques - it is also about handling data files and maintaining consistency across interview transcripts, to ensure their

security and readability. The UK Data Archive and the Finnish Social Science Data Archive have identified a number of important file-handling procedures:

- Experiment with anonymisation by first processing a couple of files.
- Make a copy of the unanonymised file and put it in a secure location. This way, possible errors in anonymisation can still be fixed.
- Begin anonymising on the copy. Files should be named clearly so that the anonymised version can be identified.
- Use specific characters, such as [...] for anonymisation to help keep track of what has been changed or not. Do not use text styling, as these changes may disappear.
- Utilise ‘find & replace’ to change names to their aliases. Be careful with the “replace all” command, since names can disappear as part of other words as well.
- Use consistent procedures within a single file and throughout the project.
- Document the anonymisation process. Keep an anonymisation log of all replacements, aggregations or removals made, in a separate place from anonymised data files.
- When anonymisation is finished, erase original files and lists of aliases. Review the background material relating to the data because they may contain identifiers (i.e. research participants’ contact information, paper questionnaires, etc.).

Note that there are, so far, few tools available for anonymising qualitative data. The UK Data Archive has developed an [add-on tool](#) based on Microsoft Word macros for helping with the anonymisation of qualitative data. It does not in itself anonymise the data, but helps in identifying potentially disclosive information to be removed or pseudonymized. It finds and highlights numbers and capital letters, which often include disclosive information.

5. CONCLUSION

This Guide has focused on the anonymisation of a specific type of qualitative data -- interview transcripts. While full anonymisation is a promise that is hard, if not impossible to keep when it comes to qualitative data, we have argued and demonstrated that it is always better to consider anonymisation together with informed consent and access control. Depending on the sensitivity of the data and the measures in place, one can apply a stronger or weaker level of anonymisation, and therefore retain more or less information with respect to reuse potential. While anonymisation techniques are rather straightforward, the main challenge consists in assessing what information to keep and what to change or remove, while maximising reuse potential. This is often the result of negotiations within the team as each project needs to be assessed on a case-by-case basis. This Guide has provided a framework and advice for helping research teams to take the best decisions, and more generally has aimed to show that qualitative data anonymisation is feasible and crucial in the context of open data.

6. RECOMMENDATIONS

To conclude, we present a set of best practices to favour:

Recommendation 1 – Only collect personal data if you really need to. The collection of personal data is subject to constraining legal obligations. The more such data you have, the more complex the anonymisation procedure may be.

Recommendation 2 – Get proper consent to handle personal data. It is important to design information and informed consent in a way that takes into account the disclosure (sharing) of data to third parties in compliance with data protection laws.

Recommendation 3 – Never promise full anonymisation. Remember that whatever you promise research participants when collecting data is a promise you have to keep, and that depending on what you said you may limit the use of your data or you may even be obliged to destroy them. Think carefully about the wording of your informed consent, be it oral or written, and do not hesitate to consult our FORS Guide on the topic ([Kruegel, 2019](#)).

Recommendation 4 – Do not limit anonymisation to direct identifiers. It takes more than simply removing direct identifiers, such as participants' names, for data to be anonymised.

Recommendation 5 – Do not confuse anonymisation with pseudonymisation. The consequences are very different, since unlike pseudonymized data, anonymised data are not subject to data protection laws.

Recommendation 6 – Do not take anonymisation too lightly. The anonymisation of data is not an exact science, and you must carefully evaluate the risks of re-identification of research participants, as well as of anyone else who was mentioned in the interview, in order to be able to take the appropriate measures.

Recommendation 7 – Do not over-anonymise. It is better to add other layers of protection, such as informed consent and access control, than to remove too much information to the point that the data become useless for future reuse.

Recommendation 8 – Plan ahead for anonymisation. You need to consider anonymisation as early as possible in the research project, ideally already at the time of the submission of the research proposal and data management plan. Keep in mind that you can request financial resources for anonymisation in grant proposals to the SNSF. Think carefully about when anonymisation should take place during the project, and assign responsibilities to ensure that someone keeps an eye on it and makes sure it is done at the right time.

Recommendation 9 – Develop an anonymisation strategy. The anonymisation strategy should be set up at the start of the project and should include a reflection on the data you need to collect, to avoid collecting personal data that are not necessary. Review the strategy as the project evolves, which includes setting up file handling rules and identifying mandatory and possible information to be anonymised.

Recommendation 10 – Choose the right anonymisation techniques. Assess potentially identifying information and the corresponding risks. Make the best choices with the aim of both protecting research respondents and retaining as much value as possible in the data.

7. Further readings and useful web links

If you are interested in general information about the anonymisation of qualitative data, you might check the CESSDA Data Management Expert Guide: <https://www.cessda.eu/Training/Training-Resources/Library/Data-Management-Expert->

[Guide](https://ukdataservice.ac.uk/learning-hub/research-data-management/). Furthermore, there is more information on the UK Data Archive website: <https://ukdataservice.ac.uk/learning-hub/research-data-management/>

If you are looking for practical data management guidelines when it comes to handling personal data and anonymising qualitative data, see the Finnish Social Science Data Archive (FSD) Data Management Guidelines: <https://www.fsd.tuni.fi/en/services/data-management-guidelines/anonymisation-and-identifiers/#anonymising-qualitative-data>

Also, you can find practical information and tutorials about data anonymisation in several previous FORS webinars on the topic at: <https://forscenter.ch/data-management-webinar-series/>

REFERENCES

- Bendjaballah, S. Cadorel, S., Fromont, E., Garcia, G., Groshens, E., Juillard, E. (2018). Anonymat et confidentialité des données : l'expérience de beQuali. *La diffusion numérique des données en SHS* (pp.207-221). Aix-en-Provence, FR : Presses universitaires de Provence. HAL Id: hal-02873570
- Coulmont (2011). *Sociologie des prénoms*. Paris, FR : La Découverte, series : « Repères »
- Diaz, P. (2022). *Data protection: legal considerations for research in Switzerland*. FORS Guide No. 17, Version 1.0. Lausanne: Swiss Center of Expertise in the Social Sciences FORS. <https://doi.org/10.24449/FG-2022-00017>
- DuBois, J. M., Strait, M., & Walsh, H. (2018). Is it time to share qualitative research data? *Qualitative Psychology*, 5(3), 380-393. <https://doi.org/10.1037/qap0000076>
- Erard, F., Heusghem, M., Parisato, C. (2023). Recherche biomédicale et Open Data. *Jusletter of 30 January 2023*. doi: 10.38023/c70abf2c-57aa-45ad-8b9a-390edc7a7f04
- Federal Act on Research involving Human Beings (Human Research Act, HRA) of 30 September 2011 (2011). <https://www.fedlex.admin.ch/eli/cc/2013/617/en>
- Kruegel, S. (2019). *The informed consent as legal and ethical basis of research data production*. FORS Guide No. 05, Version 1.0. Lausanne: Swiss Centre of Expertise in the Social Sciences FORS. <https://doi.org/10.24449/FG-2019-00005>
- Müller, B. (2006). Archives orales et entretiens ethnographiques. Un débat entre Florence Descamps et Florence Weber, animé par Bertrand Müller. *Genèses*, 62, 93-109. <https://doi.org/10.3917/gen.062.109>
- Rioufreyt, T. (2019). Réanalyser des enquêtes qualitatives à l'aide de CAQDAS. *Bulletin of Sociological Methodology/Bulletin de Méthodologie Sociologique*, 143(1), 77-106. <https://doi.org/10.1177/0759106319852902>
- Saunders, B., Kitzinger, J., & Kitzinger, C. (2015). Anonymising interview data: challenges and compromise in practice. *Qualitative Research*, 15(5), 616-632. <https://doi.org/10.1177/1468794114550439>
- Stam, A., & Kleiner, B. (2020). *Data anonymisation: legal, ethical, and strategic considerations*. FORS Guide No. 11, Version 1.1 (last update January 2022) Lausanne: Swiss Centre of Expertise in the Social Sciences FORS. <https://doi.org/10.24449/FG-2020-00011>

Svalastog, A-L., Eriksson, S. (2010). You can use my name; you don't have to steal my story - a critique of anonymity in indigenous studies. *Developing World Bioethics*, 10(2)104-110. <https://doi.org/10.1111/j.1471-8847.2010.00276.x>

Swiss Anthropological Association (SAA) (2021, November). *Open Science and Data Management in Anthropological Research*. Position Paper of the Swiss Anthropological Association (SAA) adopted at the SAA Annual Meeting, Held online. Retrieved March 9, 2023 from https://www.sagw.ch/fileadmin/redaktion_seg-sse/Position_paper_on_Open_Science_Data_Management_and_Ethics_in_Anthropological_Research/SAA_Position_paper_Open_Science_2021-11-4.pdf

Tenevic, L., & Weber, F. (1992). La délégation du travail de terrain en sociologie qualitative. *Genèses*, 8, 132-145. doi:10.3406/genes.1992.1125