FORS GUIDES
to survey methods
and data management

Data anonymisation: legal, ethical, and strategic considerations

Alexandra Stam[1] and Brian Kleiner[1]

[1]FORS

FORS Guide No. 11, Version 1.0

June 2020

Abstract:

Drawing on the Swiss context, this Guide illustrates key considerations in crafting a coherent anonymization strategy that is compliant with legal requirements and ethical norms. It shows how to strike the right balance between anonymization and other data protection measures, namely informed consent and access control, given the nature of a project's data and their anticipated use.

Keywords: anonymization, pseudonymization, data protection, research ethics

How to cite:

Stam, A., & Kleiner, B. (2020). *Data anonymization: legal, ethical, and strategic considerations*. FORS Guide No. 11, Version 1.0. Lausanne: Swiss Centre of Expertise in the Social Sciences FORS. doi:10.24449/FG-2020-00011

The FORS Guides to survey methods and data management

The FORS Guides offer support to researchers and students in 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. The FORS Guides go beyond the documentation of specific surveys or data management tools and address general topics of survey methodology. 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.

Editor:

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

1. INTRODUCTION

With increasing movement towards open data in science, data privacy has become an important and unavoidable consideration for researchers. Facilitated production and sharing of data have coincided with the reinforcement of data protection policies, with researchers confronted with a dilemma – how to protect their study participants while at the same time make their data open and available to others? Anonymisation represents a principal solution for complying with data protection requirements. Drawing on the Swiss context, this guide will address anonymization from legal, ethical, and strategic points of view[1]. The first part will present anonymization in the context of open science – what is it and why does it come into play? The second part will address anonymization from legal and ethical standpoints – what does the Swiss legal framework say about data protection and anonymization, and what are the relevant ethical considerations? The last part will provide some general guidance for formulating a coherent anonymization strategy that strikes the right balance between openness and protection. Our approach to this relies primarily on anonymisation, but ties in as well other important measures for such a strategy – specifically informed consent and access controls.

2. ABOUT ANONYMISATION

Anonymisation is a key data management practice directly related to data sharing, providing for the protection of the privacy of research participants. It consists in rendering individuals unidentifiable by removing/altering information in the data. This could be information that allows direct identification (e.g., a person's name) or pieces of information that in combination may lead to a person's identification (e.g., date of birth plus city plus gender). If applied properly, it may satisfy data protection requirements when it comes to data sharing, since anonymised data are no longer considered "personal" and therefore do not fall under the scope of data protection acts. In practice, however, full anonymisation is hard to guarantee and is often confused with other concepts, such as pseudonymisation, which retains particular legal obligations. It is therefore crucial to understand what exactly is meant by anonymisation and related concepts.

2.1 RELEVANT DEFINITIONS

We define anonymisation here 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, in the sense that it involves a **permanent action** (anonymisation cannot be reversed) and a **strong protection threshold**: re-identification should no longer be possible, or only with very intensive effort by an attacker. The second part is important - while in principle anonymisation should make re-identification impossible, studies have shown that full anonymisation is difficult to guarantee (see for example Angiuli, Blitzstein, & Waldo, 2015; Narayanan & Shmatikov, 2008; Rocher, Hendrickx, & de

[1] This Guide is the first of a series of three. While laying the theoretical basis for setting up an anonymisation strategy, the two other Guides (forthcoming) will provide more applied and technical considerations for the anonymisation of quantitative and qualitative data.

Montjoye, 2019; Tanghe & Gibert, 2017), in particular with the development of new technologies and techniques that allow for advanced computing and data linkage.

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"*.[2] Data remain pseudonymous as long as the original identifying information is somehow kept by the researcher. Unlike anonymised data, pseudonymised data remain "personal" for the holder of the related identifying keys and are therefore subject to legal obligations (see section 2.2). Researchers often confuse data anonymisation and pseudonymisation, believing that their data have been rendered anonymous while they still hold the keys, previous versions, or related documents (e.g., contact information) that allow them to re-identify participants from the research project. On the other hand, if researchers share pseudonymised data without the related identifying keys, then those data are considered anonymous for the recipients. For what follows, we consider anonymisation from an end-user perspective, independently of whether or not the data producer still holds the identification keys.

2.2 WHY ANONYMISE DATA?

The anonymisation of research data generally comes into play when sharing data with the wider research community, including colleagues and research partners. Drivers for anonymisation may be ideological, legal, or ethical. On an ideological level, research is increasingly embedded within the wider open science culture, which encourages the sharing of research data. There are a number of reasons for this, including the need for greater transparency and replication, the fact that large amounts of existing data remain un(der)exploited, and the reaffirmation that publicly funded research should belong to the scientific community. These arguments are supported by many researchers and publishers, but above all by funders. More and more funding agencies, including the Swiss National Science Foundation (SNSF), require that data used in publications be made available. Since October 2017, researchers applying for SNSF funding must submit a Data Management Plan (DMP)[3] along with their grant proposal, in which they have to state where and how the data used in publications will be made available, or else to justify why this will not be the case. This includes a description of the practices that will be implemented to this end, including data anonymisation. For more detailed information on how to draft a DMP, see our FORS Guide on the topic (Diaz & Stam, 2019). Academic journals also increasingly require that the data underlying publications be made accessible. Within the open science framework, anonymisation provides a way of complying with the various requirements and ideological values by rendering data sharable.

Anonymisation may also be triggered by legal reasons, as a way of facilitating data handling beyond a project. By continuing to hold non-anonymised personal data, researchers are legally bound and face a number of legal obligations (see section 2.0), for example with respect to ensuring data security and providing information to research participants about their data. Beyond enabling data sharing, it may also be in researchers' interest to anonymise their data in order to keep them in the long run. Indeed, data protection acts require that researchers erase personal data by the end of the research project. Furthermore, even if consent has been received to

[2] From a guide on anonymisation from FSD, the Finnish Social Science Data Archive: https://www.fsd.tuni.fi/aineistonhallinta/en/anonymisation-and-identifiers.html

[3] http://www.snf.ch/en/theSNSF/research-policies/open_research_data/Pages/data-management-plan-dmp-guidelines-for-researchers.aspx

process and keep personal data beyond the project, data need to be treated according to high security standards, and research participants will have ongoing rights over their data, placing a significant burden on the data producers. Anonymisation may also allow for the sharing of previous research data for which consent had not been obtained. Data anonymisation may therefore become an interesting alternative to data destruction or legal responsibility after the completion of a project.

Finally, anonymisation may be driven by ethical reasons, in order to ensure that research participants are protected from any harm that might arise from disclosure of their personal information. Indeed, it may also be employed by researchers even if participants have consented to their non-anonymised data being made public. There may be situations whereby research participants would want to be named (for example, as part of testimonies), but which would involve significant risks that researchers could not take ethically. Anonymisation may also be used as a way of gaining access to particular populations, by guaranteeing participants privacy protection during and beyond the project. For some participants, anonymisation may be a condition for their taking part in a study, especially if the subject matter or research context is sensitive. The next sections address more in detail the legal and ethical considerations regarding anonymisation.

3. LEGAL AND ETHICAL CONSIDERATIONS

3.1 SCOPE OF DATA PROTECTION

General

The processing of personal and sensitive data is subject to legal specifications. Data protection acts do not apply to:

- anonymised data that cannot be linked to a living person, and
- personal data that are processed exclusively for personal use and are not disclosed to third parties.

Within the research context, the following rules apply:

1. If research data contain no personal and sensitive data, then they can be kept and shared without restrictions.
2. Researchers may collect personal (sensitive) data without consent as part of a given research project as long as these are needed, but published research results must be anonymised, and any personal data must be destroyed as soon as the project for which they were gathered has ended. However, should researchers want to keep or share the data after the project, then these need to be anonymised in such a way that individuals can no longer be recognised.
3. Personal and sensitive data may only be kept or shared with appropriate informed consent and when needed for research purposes.

Swiss legislation

Treatment of personal and sensitive data is subject to specific legal obligations. In Switzerland, privately funded researchers and those working for federal institutions (like federal institutes of technology) are subject to the Federal Act on Data Protection (FADP)[4]. Researchers affiliated with cantonal institutions (such as universities, universities of applied sciences, cantonal university hospitals, etc.) must follow cantonal data protection laws. For example, in the canton of Vaud the 'Loi cantonale vaudoise sur la protection des données personnelles' (LPRD)[5] applies. Information for the other cantons can be found on the website of the Federal Data Protection and Information Commissioner[6]. In the case of a project involving collaboration between cantonal and federal institutions, both laws apply. Note that the FADP is currently under revision, and it is expected to be released in the course of 2020[7].

In addition to these federal and cantonal laws, there is also domain-specific legislation, which specifies the application of the national laws within particular domains. This is, for example, the case with the Federal act on research involving human beings (Human Research Act, HRA)[8], which applies to research concerning human diseases and the structure and function of the human body. While national laws set up general principles, domain-specific legislation provides definitions and more applied guidance.

European legislation

At the European level, the General Data Protection Regulation (GDPR)[9] applies to all researchers based in the EU and collecting personal data on people anywhere in the world, as well as researchers based outside the EU collecting data about EU citizens. Note that while the GDPR sets the general framework, the national data protection laws specify the applications. Thus, researchers in Switzerland collecting data from European citizens are subject to the GDPR and need to comply as well with the various applicable national data protection laws. There may be differences between the various European national legislations. In the case of a study involving research at the same time in Switzerland and Europe, both the Swiss and applicable European national laws need to be considered. In the case of conflicting requirements, the strictest application must be chosen. To find out about the different data protection laws worldwide, see the United Nations Conference on Trade and Development[10]**.** Keep in mind that the forthcoming revised Federal Act on Data protection will be aligned with the GDPR.

3.2 LEGAL DEFINITIONS

The processing of personal and sensitive data is subject to data protection legislation. It is therefore important to define personal and sensitive data from a legal point of view, as well as to consider how the legal framework treats anonymisation and pseudonymisation.

[4] https://www.admin.ch/opc/en/classified-compilation/19920153/index.html
[5] https://prestations.vd.ch/pub/blv-publication/actes/consolide/172.65?key=1543934892528&id=cf9df545-13f7-4106-a95b-9b3ab8fa8b01
[6] https://www.edoeb.admin.ch/edoeb/en/home/the-fdpic/links/data-protection---switzerland.html
[7] For the main expected changes, see https://www.pwc.ch/en/insights/fs/data-protection-switzerland.html. It is expected that the cantonal laws will be similar to the revised federal legislation on data protection.
[8] https://www.admin.ch/opc/en/classified-compilation/20061313/index.html
[9] https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A32016R0679
[10] https://unctad.org/en/Pages/DTL/STI_and_ICTs/ICT4D-Legislation/eCom-Data-Protection-Laws.aspx

Personal and sensitive data

Article 3(a) of the FADP defines **personal data** as "all information relating to an identified or identifiable person". The FADP further defines **sensitive personal data** (art. 3(c)) as personal data on: 1) religious, ideological, political or trade union-related views or activities, 2) health, the intimate sphere or racial origin, 3) social security measures, or 4) administrative or criminal proceedings and sanctions. Within the GDPR, sensitive data are known as "special categories of data". In addition to the categories listed in the Swiss law, the GDPR also includes biometric and genetic data (see article 9). The processing of special categories of data is prohibited under the GDPR, with various exceptions, including scientific research.

Within Swiss law one must comply with a set of principles when processing personal data, including only collecting personal data that are strictly needed (principle of proportionality). Others include good faith, recognizability, purpose, and accuracy. For more information, see the University of Lausanne's webpage on personal or sensitive data[11].

The processing and holding of personal and sensitive data gives rise to a number of legal obligations, in particular the obligation to protect personal data against unauthorized processing through appropriate organizational and technical means (article 7, FADP), as well as the obligation to provide information to research participants on the purpose of the processing, and on the data recipients in case of data sharing. Furthermore, research participants have rights over their personal data, for as long as these are kept (article 8, FADP). Any person may request information about his or her data, which needs to be provided in writing and at no cost.

Anonymisation and pseudonymisation

While the FADP does not explicitly address anonymisation or pseudonymisation (coding), the Federal Act on research involving human beings provides some guidance. As noted in the article 25 of the HRA, "*For the anonymisation of biological material and health-related personal data, all items which, when combined, would enable the data subject to be identified without disproportionate effort, must be irreversibly masked or deleted.*"

With respect to pseudonymisation, the HRA notes the following (art. 26): *"Biological material and health-related personal data are considered to be correctly coded…if, from the perspective of a person who lacks access to the key, they are to be characterised as anonymised".* The key must be stored separately from the material or data collection and in accordance with the principles of Article 5 paragraph 1, by a person to be designated in the application who is not involved in the research project.

Identifying information

Article 25 of the HRA lists the identifying information that needs to be "masked or deleted", specifically the name, address, date of birth, unique identification numbers, as well as "all items which, when combined, would enable the data subject to be identified without disproportionate effort". While it is relatively straightforward to remove unique direct identifiers such as full names, phone numbers, and social security numbers, it may be more difficult to determine which other elements when combined might lead to a person's identification (i.e., "indirect identifiers").

[11] https://www.unil.ch/openscience/en/home/menuinst/open-research-data/conformite--exigences/donnees-personnelles--sensibles.html

3.3 ETHICAL CONSIDERATIONS

Anonymisation provides a good solution for data sharing while respecting legal obligations, but it is not foolproof. The HRA recognizes implicitly the impossibility of total anonymization with the qualification that individuals should not be identifiable (when combining items) "without disproportionate effort". Indeed, full anonymization is difficult to guarantee, especially with the development of powerful analytical tools and increased access to data from multiple sources that could potentially be linked with research data and lead to a person's identification[12]. This implies that in some cases it may be possible, with intensive effort, to identify people in anonymized data. The legal responsibility of the researcher thus is to ensure that measures are taken to make such identification at least *highly difficult* (possibly by adding other measures, see 3.2). In this sense, the law factors in an acceptable level of risk to research respondents regarding their personal data.

Moreover, it is in this grey zone where ethical considerations come into play, since the law does not quantify precisely what is meant by "disproportionate". While researchers must abide by the legal requirements, they have an additional *ethical obligation* to assess what they consider to be disproportionate in a way that reduces the risk of identification to a minimum. Related to this but separate, there is the ethical obligation to ensure that study participants are sufficiently protected from *harm*. In this sense, harm is the operative notion that allows a researcher to assess the acceptable level of risk within the grey zone permitted by the law. Harm can be physical, psychological, social, economic, or legal.

Keep in mind that "identification" is not equivalent to "harm" — an individual could be identified without risk of harm, for example, if it came out that her car was the color green, or that she prefers Migros to Coop. On the other hand, she could be harmed if it were discovered that she had a criminal past or was a member of a stigmatized group. So from an ethical perspective, researchers must not only minimize to an acceptable level the risk of identification, but at the same time, in determining what is "acceptable", they must assess the risk of potential harm to their respondents if their identities were to be disclosed.

The importance of this assessment of risk of identification and harm is related to the potential *utility* of anonymized data, since in general the more information that is removed from data, the less will be their subsequent analytic value. Researchers must strike the right balance between retaining analytic value and reduction of risk to participants. With some room afforded by the law in the grey zone of "disproportionate effort", for reasons of utility the extent of anonymization should be adjusted to the level of sensitivity of data and potential for harm, thus respecting both legal and ethical injunctions. The next section addresses the considerations for striking this balance and developing a meaningful and appropriate anonymization strategy.

For more detailed information on ethical issues in relation to open research data, see our FORS Guide on the topic (Diaz, 2019).

[12] For example, in 2008 an anonymised Netflix dataset of film ratings was de-anonymized by comparing the ratings with public scores on the Internet Movie Database (Narayanan and Shmatikov, 2008).

4. DEVELOPING AN ANONYMIZATION STRATEGY

A strictly legal application of anonymization therefore needs to be handled in parallel with wider ethical considerations, in particular an assessment of the risks of harm from potential disclosure, as well as a consideration of the effects of anonymization on the potential utility of the data. Moreover, anonymization is not the only tool available to researchers for protecting respondents from harm, respecting the law, and retaining analytic utility – we will show in this section how additional measures are available toward these ends and how these can be incorporated into a larger anonymization strategy. This includes informed consent and data access controls.

4.1 ASKING THE RIGHT QUESTIONS

In developing an effective anonymization strategy, you must ask yourself the following questions regarding your project and your data:

- What is the nature and type of the personal data to anonymise? How difficult will it be to adequately anonymise the data?

Some data are more difficult to anonymise than others, or else require a greater investment of effort. For example, it is often the case that transcribed interview data can be hard to fully anonymise or need more time to remove all potentially identifying information.

- How sensitive are the data? What harm might be caused to respondents if they are identified?

You should assess what harm might be caused to respondents if they were identified. This will determine how your strategy should be shaped to address the relevant legal and ethical considerations.

- Who will be the future users of the data? Will usage be limited to researchers? What are the chances of improper use?

The set of future users is important, since the more people who have access to the data, the greater the chances of identification. Also, the types of users should also be considered – limiting access to authenticated researchers should reduce 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?

It is important to consider how your data might be used in the future, including how much information and detail should be retained in order to address particular uses. This will help determine the appropriate level of anonymization.

- What should be promised to respondents regarding the future use of their data? For cases where the anonymization strategy is decided toward the end of a project, what was promised in the consent form?

If you are developing your anonymization strategy at the beginning of your project, then you will be able to factor this into what you promise your respondents and how you formulate an informed consent form. If you devise your strategy at the end of your project, then you will be constrained by whatever you have promised. For more detailed information on how to draft a consent form, see our FORS Guide on informed consent (Kruegel, 2019).

In general, a good practice in developing an anonymisation strategy consists in first conducting a risk assessment, taking into consideration the type of harm that may occur (e.g., social harm, psychological harm, physical harm), its intensity (small, medium, strong), and the likelihood of identity disclosure. Likelihood of disclosure is strongly linked to the characteristics of the data (type of data, research methodology, sampling methods, research domain, nature of the topic, age of the data, etc.).

4.2 COMBINING THREE KEY ELEMENTS: LEVEL OF ANONYMIZATION, INFORMED CONSENT, AND ACCESS CONTROLS

Given the considerations from the previous section, you should then calibrate three elements that will allow you find the appropriate balance between data openness, utility, and protection – these are 1) level of anonymization, 2) informed consent, 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 anonymization, stronger promises regarding anonymization and access, and stricter controls on access.

For less sensitive data, calibrating between the three elements means that if you apply more for one element, then you may be able to apply less for another. To take an example, if you determine that your data cannot easily be anonymized, or that you do not have sufficient resources to do so, then you might do less anonymization but apply stricter controls on access. This might mean limiting the pool of possible users or requiring your own permission before granting access. Repositories such as FORSbase[13] offer possibilities for limiting data access to specific groups of users or for particular purposes (e.g., for research, but not for teaching), as well as requiring permission. This is fortified at FORS by requiring that end-users sign a user-license by which they are legally bound not to try to identify specific individuals. To take another example, if you consider that high potential utility is a priority, then you can employ an informed consent that does not promise full anonymity, of course only if the risk of harm to respondents is minimal.

In these ways you can develop an anonymisation strategy that is well-suited to your specific project and data. Keep in mind that every project is different and there is no one-size-fits-all solution for anonymization. Most important is to ask the right questions and put into place a strategy that minimizes risk to respondents while maximizing data openness and utility.

5. TWO EXAMPLES OF AN ANONYMISATION STRATEGY

5.1 CASE STUDY 1: ANONYMISATION STRATEGY FOR THE SWISS DATA FROM THE EUROPEAN SOCIAL SURVEY (ESS) DISTRIBUTED BY FORS

Topic and research methodology

The European Social Survey (ESS) is a cross-national survey that has been conducted in around 30 European countries every two years since 2002. In each country, a minimum of 1,500 respondents take part in a one-hour face-to-face interview. Switzerland has participated in all rounds since the very beginning. The ESS measures values, attitudes and behavioural patterns

[13] https://forsbase.unil.ch/

of the populations of European countries. The respondents are drawn from a probabilistic sample representing the countries' population aged 15 and above.

Respondents are provided information in an advance letter on the following:

- Who the Data Controller and Data Processors are
- The purpose for which the data is being collected
- Where the data will be stored and for how long
- A note about sensitive information that will be collected, e.g., life events, social and political attitudes
- A note about the voluntary nature of the research and the right not to answer specific questions
- Information about data collected that is not part of the interview (e.g., contact form data, neighbourhood characteristics)
- Name and contact details for the ESS ERIC Data Protection Officer

Participation in the survey is considered as informed consent.

After completion of the fieldwork, FORS reviews and verifies the data and processes them, so that they can be integrated into the central data archive of the ESS. With respect to anonymization, there are general guidelines since 2018 provided by the ESS central coordinating body, but their implementation in the details is carried out by the Swiss ESS team, which has followed the same anonymization principles since the beginning of the ESS. It makes sense that the national teams do this work, since they have the best access to expertise on their national population's size, composition and demographic variation, which define disclosure risks. In accordance with data protection regulations in participating countries, only anonymised data are made publicly available to users. Before depositing data to the ESS Archive, each national team is responsible for checking their data with confidentiality requirements in mind. National teams are asked to confirm in their National Technical Summary that all data that will be made publicly available to users have been anonymised in accordance with national and EU regulations, including the General Data Protection Regulation (GDPR). Once a country's data have been published, the survey agency and the Swiss National team are required to delete all names and addresses of respondents, as well as the key that links the serial number to the names and addresses.

Anonymised ESS datasets are freely available to researchers after registration on FORSbase[14] and agreement to the data use conditions. The Swiss ESS metadata are also available on the FORS – De Visu Server[15]. This server provides exclusive access to additional, country-specific questions surveyed in Switzerland, and to the German and French language versions. The overall dataset is available in English on the FORSbase data catalogue.

Risk evaluation

The ESS collects personal data from participants, including their opinions, attitudes, and living conditions. However, the data from the ESS are not especially sensitive, and in general identity

[14] https://forsbase.unil.ch/
[15] https://devisu.forscenter.ch/index.php/catalog/central/about

disclosure would not lead to harm to respondents. Direct identifiers are permanently removed from the file kept by the national team. In addition, direct identifiers (such as name and phone number) are removed from the public data file and certain indirect identifiers (such as profession and geographical location) are recoded, and so the risk of identification of individual participants is very low. On the other hand, it would not be impossible with intensive effort to combine certain variables in the file to identify certain individuals (i.e., if they accumulate several rare characteristics), and so a total anonymization is excluded. This means that in addition to anonymizing the data, other measures are needed to reduce risk and potential harm to participants, notably access controls and data use conditions.

Sharing the Swiss ESS data at FORS

The FORS data archive works with the Swiss team of the ESS to ensure that the Swiss data are made available through FORSbase. The following measures are in place as part of the Swiss team's ESS anonymization and data protection strategy for the Swiss data distributed at FORS:

1. Only authenticated users may obtain access to the data file.
2. Users must be affiliated with a research institution.
3. Users must describe why they need to access the data.
4. Users are required to sign a user license that binds them to respect a set of data use conditions, in particular not to try to identify individual respondents, not to share the file with third parties, and to properly cite the data in articles or other forms of publication.
5. Users must delete the dataset at the end of the license period.

In sum, given the fact that data can be anonymized to a high degree, and that the risk of identification and harm is relatively small, the data are available under default conditions at FORS – user ID/authentication, user license, and a justification and description of their proposed project/analyses.

The anonymization carried out on the Swiss ESS data allows for retaining a great deal of utility. Nonetheless, removal of certain variables does have an effect on the potential usefulness of the data. In order to maximise the utility of the data, certain of the removed variables are available at FORS under even more restrictive conditions. A file with additional variables that might increase the risk of identification (i.e., by crossing variables) is available upon special request to the Swiss ESS team, who evaluate whether to grant access on a case-by-case basis.

5.2 CASE STUDY 2: BASED ON THE ARTICLE "ANONYMISING INTERVIEW DATA: CHALLENGES AND COMPROMISE IN PRACTICE" (SAUNDERS, B; KITZINGER, J; & KITZINGER C., 2015)

Topic and research methodology

A research team has collected data between 2010 and 2013 on brain injury in the form of in-depth qualitative interviews with family members of people in vegetative and minimally conscious states. In some cases, several people of the same family were interviewed separately. Participants filled in a consent form including a range of permissions, from giving consent for only the project team to access the data to archiving the data in a repository. The research team promised participants to keep their identities hidden as far as possible by changing their names and those

of the people they mention, as well as other identifying details. The research team also informed participants about the limits to the anonymity they could offer, as well as the challenges posed to maintaining anonymity.

Risk evaluation

The collected data are not only personal but also sensitive, since they relate to the topic of health. Identify disclosure might have consequences not only for the interviewees but for others as well, since the interviewees may release sensitive information about the brain injured people themselves, healthcare professionals, and the concerned institutions (specific hospitals, care centers, etc.). Risks of identification are high, since the population of brain-injured people is small, even at the country level. Furthermore, some of the stories of the injured may be public, either because of media coverage, or possibly law cases. Linkage with other data sources, including the social media sources of the research participants, represents further important threats to confidentiality. Depending on the content of the narratives, harm could range from minor to very significant.

Options for sharing within a repository

At the time the article was published, the research team had not yet archived data in a repository. If FORS was consulted, we would make the following recommendations:

- Given the fact that potential for harm differs across cases, we would advise the research team to consider the narratives individually. While it is good practice to define an anonymization strategy based on the specific elements to be anonymized, the research team should assess risk on a case to case basis. In some cases, for example, family members confessed their vulnerable emotional states, stressing that they never told this to their relatives. When dealing with sensitive information, the research team must evaluate for each case how important that information is for the interpretation of the narratives. If removal of specific information would affect the quality of the narratives and potential for re-use, then decisions need to be made regarding how best to protect research participants and indirect actors (people mentioned in the interviews, institutions). If the research process allows it, it may be an option to discuss this with the research participants themselves. For instance, Saunders et al. (2015) mentioned that as they gained experience through carrying out the interviews they were able to identify possible problematic situations and discuss them on the spot with research participants. Other researchers choose to share anonymized transcripts with their respondents after the fact, even though this may be complicated if anonymization is done at a later stage in the research process.
- Rather than over-anonymize narratives and therefore lose important re-use potential, we would advise the research team to consider combining anonymization and access controls, in relation to what was promised in the informed consent. Only the research team is fully able to judge the sensitivity of the data, the associated risks of disclosure, and the re-use potential of the data. We therefore cannot state what the optimal solution would be in this case. However, from a general perspective, it would seem reasonable to deposit the interview transcripts that underwent careful anonymization, provided that respondents were made aware of potential identification risks and agreed for their data to be deposited in an archive. Furthermore, if the data cannot be sufficiently anonymized

without sacrificing too much utility, we would recommend that the research team restrict access to the data. This is possible when depositing data in established archives, such as FORSbase. Data depositors may decide on the conditions of access (e.g., for research purposes only, or only with prior approval of the data producer). Furthermore, data users not only need to register to access data, but they also need to sign a contract by which they commit not to seek identification of the research participants. For this case, we could imagine adding a few conditions in the contract, for example, how to handle extracts within publications.

Other considerations

For some complex research projects, like the one described in the article by Saunders et al. (2015), anonymization needs to be undertaken at an early stage, in order to share data within the research team. Anonymisation is therefore not always undertaken at the end stage, but may be an important part of the research process itself. Furthermore, when dealing with sensitive data, particular attention may also need to be given to individual data extracts published in journals. For more details on anonymization during the research project and within publications, we recommend you read the article by Saunders et al. (2015).

6. RECOMMENDATIONS

Recommendation 1 – Plan your anonymization strategy at the beginning of your project. Waiting until the end may severely limit your possibilities.

Recommendation 2 – Even if exhaustive anonymization is foreseen, we recommend that researchers always ask for informed consent. The consent form should be transparent and explain what will be done with the data.

Recommendation 3 – Always consider anonymisation in relation to risk of harm and legal requirements, but also in relation to other forms of protection to respondents, such as informed consent and access controls. This will allow you to optimise data openness, utility, and protection.

7. FURTHER READINGS AND USEFUL WEB LINKS

If you are interested in strategic and legal considerations for undertaking anonymisation you may consult the anonymisation decision-making framework: http://ukanon.net/wp-content/uploads/2015/05/The-Anonymisation-Decision-making-Framework.pdf

For more information on data protection legislation in Switzerland, see: https://www.lexfutura.ch/fileadmin/lexfutura.ch/Bilder/Blog/Data_Protection_Switzerland.pdf

REFERENCES

Angiuli, O., Blitzstein, J., & Waldo, J. (2015). How to De-Identify Your Data. Communications of the ACM. 58. 48-55. Doi: 10.1145/2814340

Diaz, P. (2019). *Ethics in the era of open research data: some points of reference*. FORS Guide No. 03, Version 1.0. Lausanne: Swiss Centre of Expertise in the Social Sciences FORS. doi:10.24449/FG-2018-00003

Diaz, P., & Stam, A. (2019). *How to draft a DMP from the perspective of the social sciences, using the SNSF template*. FORS Guide No. 07, Version 1.0. Lausanne: Swiss Centre of Expertise in the Social Sciences FORS. doi:10.24449/FG-2019-00007

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. doi:10.24449/FG-2019-00005

Narayanan, A., & Shmatikov, V. (2008). Robust de-anonymization of large sparse datasets. In Proceedings - 2008 IEEE Symposium on Security and Privacy, SP (pp. 111-125). doi: 10.1109/SP.2008.33

Rocher, L., Hendrickx, J. M., & de Montjoye, Y. (2019). Estimating the success of re-identifications in incomplete datasets using generative models. *Nat Commun* 10, 3069. doi: 10.1038/s41467-019-10933-3

Saunders, B., Kitzinger, J., & Kitzinger, C. (2015). Anonymising interview data: challenges and compromise in practice. Qualitative Research, 15(5), 616-632. doi: 10.1177/1468794114550439

Tanghe, H., & Gibert, P. (2017). L'enjeu de l'anonymisation à l'heure du big data. *Revue française des affaires sociales*, 79-93. doi:10.3917/rfas.174.0079