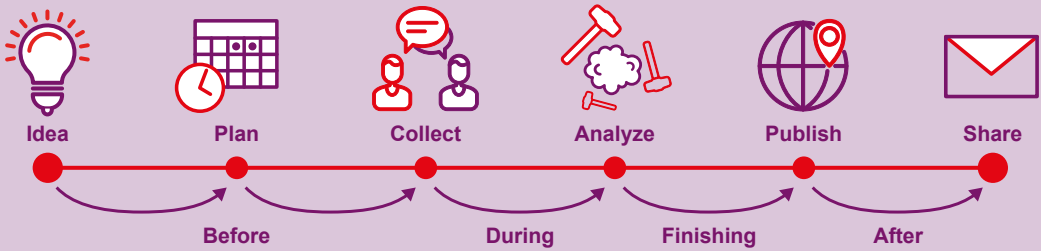
university of groningen

Data minimization and de-identification

How to protect the identity of your research data subjects



1 Before your research

- Design your project to minimize the collection of personal data.
- Think about **de-identification strategies** to apply in different stages of your research.
- Include data minimization and de-identification strategies in your Research Data Management Plan (RDMP).

2 During your research

- While doing your research you can rely on the **protocols** for de-identification and data management that you developed in the planning phase.
- Check whether your RDMP is still up to date.

3 Finishing your research

- Archive and publish your data in accordance with the FAIR guiding principles.
- Assess the sensitivity of your dataset to determine whether there are reasons to restrict access.
- Apply appropriate de-identification techniques to share your research data responsibly.

4 After your research

- Make sure you do not keep data that is redundant and clean out your working directory.

Learn more at the UG Digital Competence Centre:
rug.nl/de-identification
For questions or support: dcc@rug.nl

Table of content

Introduction

As a researcher you are responsible for protecting the privacy of your data subjects. For this reason, the principle of data minimization (GDPR art. 5 (1c)) should be one of the leading in the design of your research project. This means that you only collect personal data that is necessary for your research purposes, and de-identify your dataset once personal data is no longer needed to prevent re-identification of your data subjects.

Implementing these safeguards is especially important when you collect sensitive personal data, share data with collaborators or make data available for reuse or verification purposes.

In this guide[1] you will learn about

- Essential concepts for data minimization
- How to design your research to limit the collection of personal data
- Various de-identification techniques

[1] Keep in mind that some of the recommendations in this guide might not apply to medical research. The UMCG has its own Standard Operating Procedures (SOPs) for complying with the GDPR which overrule instructions in this document (For more information).

Essential concepts related to data minimization

Personal data

"Personal data is any information that relates to an identified or identifiable living individual. Different pieces of information, which collected together can lead to the identification of a particular person, also constitute personal data." (European commission). Some personal data are sensitive by nature and therefore require extra protection. Examples of sensitive data are provided in, but not limited to, the list of special categories of personal data, defined in the GDPR (GDPR art. 9(1))

Direct identifiers

Direct identifiers are data that make it easy to identify an individual, such as name, e-mail address, phone number, home address or IP address.

Indirect identifiers

Indirect identifiers (or: quasi identifiers) are data that do not directly identify an individual, but could, in combination with other identifiers, be unique to an individual and can therefore lead to identification. For example: Women from Groningen who drive a McLaren (car). Combined the underlined identifiers could possibly single out an individual and are therefore examples of indirect identifiers.

Granularity

Data granularity refers to the level of detail in a data structure or variable. (C3 AI). The higher the granularity in a dataset, the higher the possibility of re-identification.

Data minimization

Data minimization is one of the data protection principles that form the basis of the GDPR. It states that the processing of personal data should be "adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed" (GDPR art. 5 (1c)); Data minimization does not mean that you cannot collect personal data at all. If you can explain why you need these data for the current or specific future purposes you are allowed to collect these data.

De-identification

De-identification is the masking, manipulation or removal of personal data with the aim to make individuals in a dataset less easy to identify.

Pseudonymisation

Pseudonymization is a de-identification procedure during which personally identifiable information is replaced by an unique alias or code (pseudonym). In general, the names and/or contact details of data subjects are stored with this pseudonym in a so-called keyfile. The keyfile enables the re-identification of individuals in the dataset. Keyfiles are stored separately from the rest of the data and access should be restricted. In contrast to an anonymized dataset, a pseudonymized dataset in principle still allows for the re-identification of data subjects.

Anonymization

Anonymization is a de-identification procedure during which "personal data is altered in such a way that a data subject can no longer be identified directly or indirectly, either by the data controller alone or in collaboration with any other party." (ISO 25237:2017 Health informatics -- Pseudonymization. ISO. 2017. p. 7.). In contrast to a pseudonymized dataset, an anonymized dataset does **not** allow for the re-identification of data subjects and is therefore no longer considered personal data.

Designing your research to limit the amount of personal data

Before your research

As a researcher you are responsible for protecting the privacy of your data subjects. It is important to design your project in such a way that you minimize the collection of personal data. Creating a Research Data Management Plan (RDMP) and a project plan helps you to reflect on the kind of (personal) data you will collect and how you will handle these in different stages of your research. Below you find a list of important questions to address. For additional support, DCC data stewards are available to provide tailored advice.

What personal data do I need to answer my research questions?

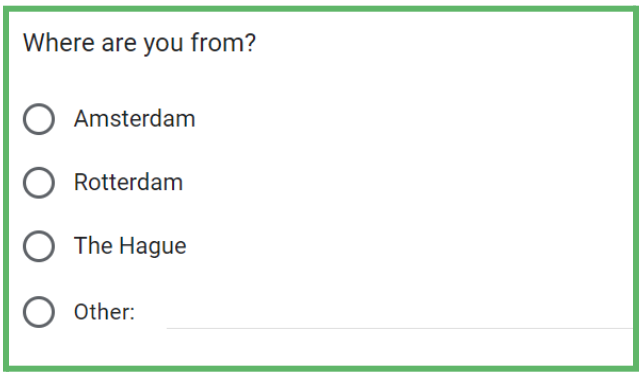
- Consider what personal data you need to answer your research questions.
  Good practice: Limit the collection of personal data to what you need for your research.
- Consider what demographics you need to describe the data subjects and which control variables you need for your analyses.
  Good practice: Be critical about conventions that might be outdated and pay special attention to special categories of personal data.
- Are there additional questions that you may want to address with your dataset? Avoid having to repeat your data collection with a new pool of participants.
  Good practice: Explain why you need these data for your current or future research purposes.

What level of granularity do I need to answer my research questions?

- Consider the type of data you need.
  Good practice: Only use a rich data collection method, if you also use this type of data for your analyses.
  - Video: Facial expressions or movement patterns.
  - Sound: speech analysis or open interviews.
  - Text: questionnaires, surveys or structured interviews.
  - Example: you do not need video recordings of interviews if you only need the transcriptions; use audio recordings instead.
- Consider the level of granularity you need.
  Good practice: Decide whether data minimization through generalization is feasible.
  - **Example 1:** Postal code:neighborhood:city:province:country.
  - **Example 2:** Birthday:birth month:birth year:age:age group.

- **Example 3:** By asking open questions ('Where are you from?'), you risk collecting more data from your subjects than you actually need. Instead you can offer options.

Figure 1: Example of open questions that could expose more information than necessary for the research purpose.



- **Further reading:**
  - LCRDM report on risk management
  - Location Privacy patterns

Am I aware of personal (meta)data that are automatically generated through my data collection method?

- Online (survey) tools such as Qualtrics sometimes automatically register IP addresses.
  Good practice: Check whether it is possible to turn off automatic IP address registration. **Example:** Qualtrics
- Video or audio files might contain a timestamp, date and depending on the method also location.
  Good practice: Check whether you can prevent the collection of these data or remove these metadata as soon as possible after collection. **Example**: Comparitech.com
- Is there a social media ID or patientID included in your dataset?
  Good practice: If you do not need this ID for current or future research (e.g. connect to other datasets), delete these IDs from your dataset. Consider pseudonymization if you do need these IDs for your research.
- Be aware that through online calendar invitations or online interviews personal data about data subjects might be visible to others.
  Good practice: set appointments in 'private' mode; share video call-links by email.

How can I make it harder to identify data subjects in my dataset?

➢ Determine whether it is necessary to keep a connection between your research data and the data subjects.
Good practice: Make a pseudonymization protocol.
- **Example 1:** Simple pseudonymization procedure with a keyfile

You can pseudonymize a basic quantitative dataset through separation of the contact information from the rest of the data. It is still possible to re-identify a specific participant in this dataset through the ID which is present in both files. Therefore, it is important to keep the information with contact information in a separate location or folder from the rest of the data (e.g., y-drive, unishare or RDMS), preferably protected by encryption. Keep track of who has access to the key-file; make sure that there are always two people who have access to the key in case something happens. Do not keep this keyfile longer than is necessary for your research (e.g., exercising data subjects' rights or informing data subjects about the research).

Figure 2: a. Keyfile with contact details

	A	B	C
1	**ID**	**Name**	**contact information**
2	20877	Anne	Anne@live.nl
3	29896	Bert	Bert@hotmail.com
4	28515	Catherine	Catherine@gmail.com
5	25827	Daphne	Daphne@outlook.com
6	25473	Earl	Earl@rug.nl
7	28050	Francis	Francis@Francis.com
8	20111	...	...

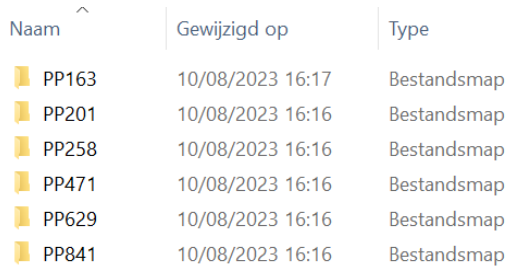
b. Datafile without direct identifiers

	A	B	C	D	E
1	**ID**	**Age**	**Q1**	**Q2**	**Q3**
2	20877	46-50	5	1	3
3	29896	41-45	1	2	3
4	28515	16-20	4	3	4
5	25827	36-40	2	2	4
6	25473	21-25	4	2	1
7	28050	36-40	4	5	2
8	20111	31-35	5	4	1

- **Example 2:** Simple pseudonymization procedure without a keyfile

It is also possible to pseudonymize your research data by connecting different datafiles through a pseudonymization ID (e.g., transcripts, survey data and other data) without using a keyfile with directly identifiable information. It might still be possible to re-identify some data subjects in your data because, in combination, your research data might single out an individual (e.g., combination of height, job occupation and location of data collection).

Figure 3: a. Folders with pseudonyms.



b. File names with pseudonyms.

- **Example 3:** Pseudonymization protocol with double coding

In some research projects, you might need to query data about your data subjects from external organizations (i.e., health information from medical files). Especially when data are sensitive, it is important to make sure that the other organization pseudonymizes the data before transfer. In this case, it is advisable to use an extra layer of pseudonymization (double coding) to make sure the other organization does not have the pseudonymization ID that is used in your research data. Contact your faculty data steward or the DCC for more information on this protocol.

Figure 4a: Keyfile with contact details.

	A	B	C	D
1	**ID_1**	**ID_2**	**Name**	**contact information**
2	802	20877	Anne	Anne@live.nl
3	627	29896	Bert	Bert@hotmail.com
4	215	28515	Catherine	Catherine@gmail.com
5	576	25827	Daphne	Daphne@outlook.com
6	980	25473	Earl	Earl@rug.nl
7	273	28050	Francis	Francis@Francis.com
8	810	20111	...	...
9				

b. Hospital data with pseudonymization ID_1

	A	B	C	D	E
1	**ID**	**days_hosp**	**symp1**	**symp2**	**symp3**
2	802	10	1	1	3
3	627	7	4	2	3
4	215	3	5	2	4
5	576	1	4	5	4
6	980	1	2	4	1
7	273	0	3	4	2
8	810	8	5	2	1

c. Data with pseudonymization ID_2

	A	B	C	D	E
1	**ID**	**Age**	**Q1**	**Q2**	**Q3**
2	20877	46-50	5	1	3
3	29896	41-45	1	2	3
4	28515	16-20	4	3	4
5	25827	36-40	2	2	4
6	25473	21-25	4	2	1
7	28050	36-40	4	5	2
8	20111	31-35	5	4	1

- Ask your faculty data steward or the DCC for tailored advice.

➢ Plan for minimizing personal data in your dataset. Draw up a procedure for removing personal data that is no longer necessary for your research. Do not delete data that you need for verification purposes.

- Consider the degree of identifiability necessary for your research.
  Good practice: Include the de-identification measures in your research protocol or data management plan (e.g., removal of email address from dataset, transcription of interviews).
- Consider whether video or audio data are still necessary for your current or future analyses or to verify your results.

Good practice: Create a destruction protocol based on the UG template for the destruction of audio and video materials. Ask the DCC for the current template and advice on this procedure.

- Be aware that consent from your participant can reveal personal information.
  Good practice: Plan to handle consent registration with care.
  - Paper consent: Scan paper consent forms; archive digitized consent forms separately from your research data and destroy the original paper forms (use UG paper containers for confidential materials or a shredder).
    - ★ If your objective is to collect anonymous data, do not ask for names and do not use pseudonymization IDs in consent forms.
    - ★ If your objective is to collect (pseudonymized) personal data, use a pseudonymization ID in consent forms to prevent direct identification.
  - Audio or video consent: Make sure the verbal consent recorded via audio or video is saved separately from your research data (e.g., experiment, interview, observation etc.); Archive the consent files in a separate location from your research data; Be aware that audio or video recordings of informed consent cannot be de-identified; use an extra layer of protection, such as encryption.
- Make sure that you do not include contact information or other personal data in the naming of your files.
  Good practice: Organize your data consistently by using a file naming strategy and good folder structure.
- Be aware that correspondence with your data subjects also contains personal data.
  Good practice: Plan to manage your correspondence with data subjects. Put the general content/message of emails in a log/protocol file and plan to remove these mails when appropriate. Plan to remove personal data that participants may provide at their own initiative and which is not necessary for the purpose of your research.

Have I considered where I store my research data?

➢ Limit data storage locations. Consider where you store (personal) data during your research. Do not keep versions of your data on devices and platforms that you no longer use.

- Use recommended UG IT solutions that fit the sensitivity of your data.
- Plan to remove your dataset from portable devices as soon as possible (e.g., audio or video recorder, USB flash drives).
- Plan to remove (personal) data from the questionnaire platform (e.g., Qualtrics), after you stored your dataset on your analysis platform.
- Check whether your storage solution already makes backups automatically. Don't create more back-ups than necessary.

Who will have access to my data?

- ➢ Plan for access management and consider who needs access to what data and when.
    - This is especially important if you work with students and student assistants, collaborators outside of the university, external parties that process personal data under your responsibility (e.g., transcription service/tool).
      Good practice: Create your data access protocol. Make sure parties only receive access to the data that they need for their role in the project and that access rights remain up to date (e.g., offboarding of team members or students). Use the UG research data policy and your faculty data management protocol as guidance. Review regularly (e.g. twice a year), whether the right people still have access to the right data.

During your research

While doing your research you can rely on the protocols for de-identification and data management that you developed in the planning phase.

Do I have to do anything before I can start collecting data? Am I following the procedures I prepared?

- ➢ Confirm whether you have included all the necessary procedures in your RDMP.
- ➢ Check your RDMP for procedures that you need to follow regarding data minimization and de-identification.
    - Remove (meta)data that are automatically generated through my data collection method.
    - Execute the pseudonymization procedure you designed.
    - Follow the procedure that you designed for the removal of personal data that is no longer necessary for your research.
    - Remove versions of your data from devices and platforms that you no longer use.
    - Manage access to your data according to your plan
- ➢ Register whether your data subjects agreed to the reuse of their personal data and by whom and for what purpose.
    - **Example 1:** Include an extra column with information about reuse in your structured dataset.
    - **Example 2:** Include an extra column with information about reuse in a data subject file with general information about your data subjects.

Is my RDMP still up to date?

- ➢ Check whether it is necessary to revise your planned procedures, because research ‘is what happens to you while you’re busy making other plans’.

Finishing your research

Your project nears its completion. It is time to prepare your data for archiving and publication in accordance with the FAIR guiding principles with the goal of making your data as open as possible and as closed as necessary. De-identification techniques enable you to share your research data with other researchers, while protecting the privacy of your data subjects.

Which data should I archive after my project ends?

- ➢ Follow the procedure in the removal protocol(s) that you designed at the beginning of your project. Add these protocol(s) to your data package, publication package or archive.
- ➢ Check whether you can further minimize your data, with two goals of archiving in mind:
  - Select and organize the data that are needed to validate your findings;
  - Select and organize the data that are potentially valuable for further research by you, your team, or fellow researchers.
- ➢ Consult the research data management policy of your faculty or institute to assess whether you comply with the data storage and sharing requirements.

How can I make my data as open as possible and as closed as necessary?

- ➢ FAIR data does not necessarily mean that your data is openly available. There can be good reasons to restrict access to your data. The objective is to have data as open as possible, and as closed and protected as necessary.
  - Experts from the UG DCC can give tailored advice.

- ➢ Consider applying a 'layered' approach to your files by scoring your files in terms of sensitivity
  - Category 1: contains no personal data
    Good practice: publish your dataset in a recognized data repository such as DataverseNL, on the condition that no other reasons for restricting access apply. Allow for reuse by adding a license (for instance a CC0 license) and use the DOI for data citation.
  - Category 2: contains well de-identified data, with pseudonymisation.
    Good practice: publish your dataset in a recognized repository such as DataverseNL, under restricted access. Determine the terms of access and use for external parties that would like to reuse your data.
  - Category 3: contains sensitive personal data.
    When your data still contains highly sensitive information, do not publish data in a data repository. Instead, archive your data in accordance with the

research data policy of your faculty or institute. The UG DCC can assist in developing a procedure for making these sensitive data available for reuse under well-defined conditions.

Good practice: determine whether it is possible to de-identify your category 3 files further while maintaining the archiving goals. Check out possible techniques to de-identify your data:

- **Removal:** Consider whether you can remove or suppress sensitive elements (e.g., patient IDs). If this is feasible your dataset can be handled as category 2 data. Be aware that this technique often affects the analytical value of the dataset.
- **Replacement**: A practice in which you replace sensitive personal data with values or codes that are not sensitive. **Examples** (UU):
    - Replace direct identifiers ('name') with a pseudonym ('X')
    - Make numerical values less precise
    - Replace identifiable text with '[redacted]'.
- **Masking or hashing**: Replace a character of a data value in your dataset with a constant symbol (for instance an 'X' or '*'). Masking is typically partial, i.e. applied only to some characters in the attribute. Example:
    - Postal code: 9746DC → 97****
- **Aggregation & Generalization**: Reduce the granularity of your dataset by generalizing variables, which makes it harder to identify individual subjects. This can be applied to both quantitative and qualitative datasets. **Examples**:
    - Address → city
    - Age → age group
- **Bottom- and top-coding (winsorizing)**: Can be applied to datasets with unique extreme values. Set a maximum or minimum and recode all higher or lower values to that minimum or maximum. Replace values above or below a certain threshold with the same standard value. **Example** (UU):
    - Top-code the variable 'income' by setting all incomes over €100.000 to €100.000. This distorts the distribution, yet leaves a large part of the data intact.
- **Perturbation or adding noise (differential privacy)**: Obscure the sensitive elements in your dataset, e.g. by removing obvious attributes and quasi-identifiers. Noise addition is usually combined with other anonymisation techniques and is mostly (but not always) applied to quantitative datasets. **Examples**:
    - Add half a standard deviation to a variable.
    - Multiplying a variable by a random number.
    - Blur photos and videos or alter voices in audio recordings.
- **Permutation**: Applied to quantitative datasets. Shuffle the attributes in a table to link some of them artificially to different data subjects. The exact distribution per attribute of the dataset is hereby retained, but identification of data subjects is made more difficult.

Further reading:

- Privacy handbook Utrecht University
- CBS Statistical disclosure control
- IAPP guide to basic data anonymization techniques
- An overview of anonymization techniques and tools from the University of Brussel.

Example: Two worked examples of making data GDPR compliant and FAIR.

How can I be sure that my data is anonymous or sufficiently de-identified?

- ➢ The University of Utrecht has created an overview of statistical techniques that could help you test the identifiability of data subjects in your dataset.
- ➢ When you publish your data on DataverseNL, the curators perform a basic assessment on the risk of re-identification before you are allowed to publish the data.
- ➢ For additional support on how to de-identify your data before publishing, DCC data stewards are available to provide tailored advice.

How do I manage who has access to my 'restricted access' data?

- ➢ Design a data access policy with terms of use and terms of access.
  - **Example:** DALC dataset (Anonymized data: GitHub & Restricted data: DataverseNL).
  - **Example:** PsyCorona (Restricted access: DataverseNL).
- ➢ Draw up your data availability statement to include in your journal article. Make sure this statement is in line with the data subjects' consent.

After your research

Although your research project is over, the research data that was generated during your project still needs to be managed. Check whether all necessary data is included in your archive, clean out your working directory, and organize the handling of requests for access.

Have I removed data that is redundant?

- ➢ Check whether all necessary data is included in your data package, publication package or archive according to the faculty research data policy.
- ➢ If you are no longer working with the data you have archived, clean out your working directories.

Recommended reading

- Risk management for research data with about people (incl. pseudonymization) (LCRDM)
- About pseudonymization and securing the pseudonymization key (LCRDM)
- Removing identifiers from human data (guide from the University of Sydney)
- De-identification training exercises by the UK Data Service (exercises using quantitative and qualitative data)
- 10 misunderstandings about anonymisation (European Data Protection Supervisor)