Data Privacy Handbook

Utrecht University | Last updated: 2023-05-31

31 mei 2023

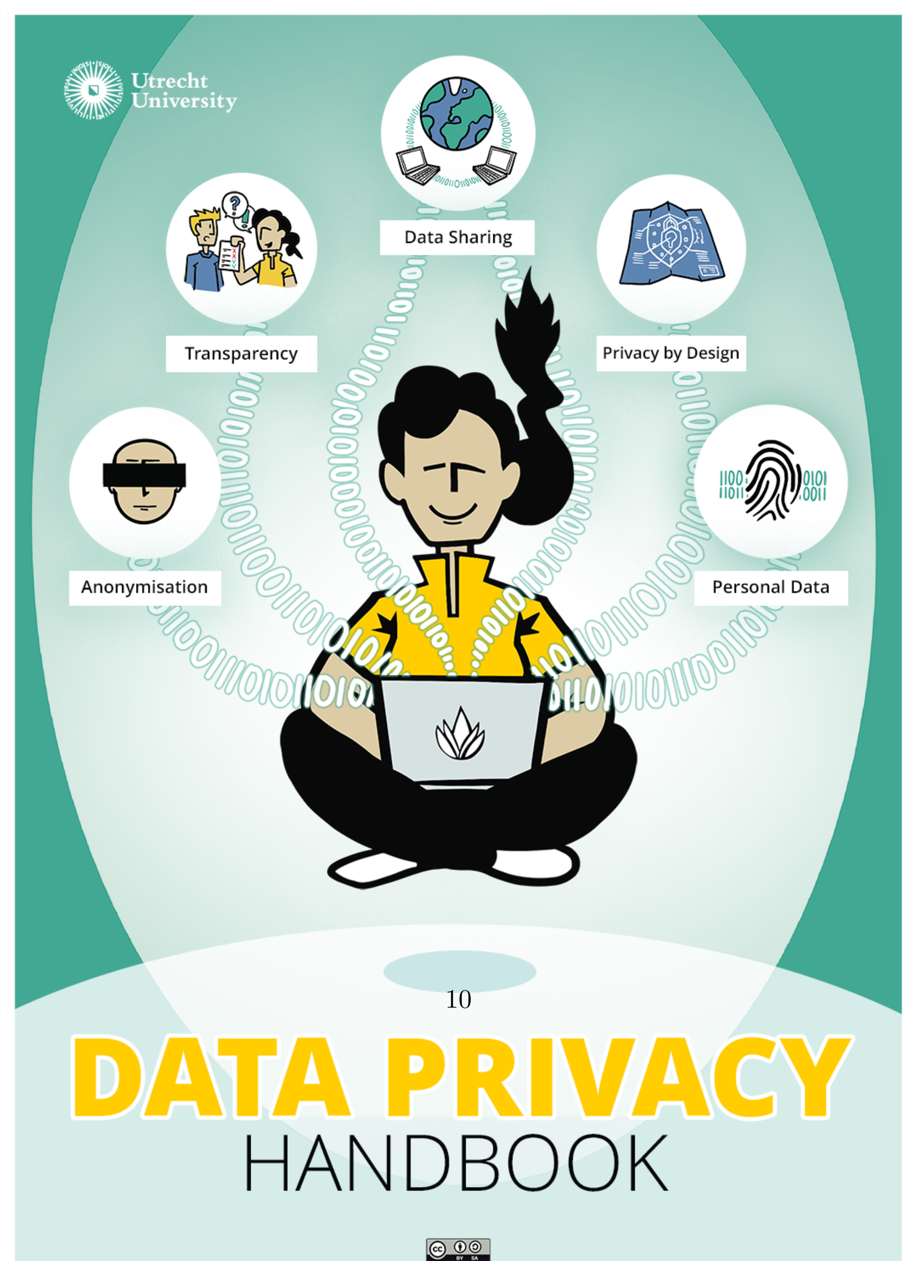
Contents

Resources 163

Intro

Chapter 1

Data Privacy Handbook



Last Handbook update: 31 May 2023

1.1 About

The Data Privacy Handbook is a practical guide on handling personal data in scientific research, primarily written for Utrecht University researchers and research support staff in the Netherlands. It is an initiative of Research Data Management Support, in collaboration with privacy and data experts at Utrecht University.

The Data Privacy Handbook consists of:

- A **knowledge base** which explains how the EU General Data Protection Regulation (GDPR, Dutch: Algemene Verordening Gegevensbescherming) applies to scientific research, including guidelines and good practices in carrying out GDPR-compliant scientific research;
- An overview of privacy-enhancing **techniques & tools** and practical guidance on their implementation;
- **Use Cases** in the form of research projects with privacy-related issues, for which a reusable solution (e.g., tool, workflow) is shared.

This is an Utrecht University (UU) community-driven, open source project. We welcome feedback and contributions of any type, please read our contributing guidelines for more information.

If you work at the University Medical Centre Utrecht (UMCU) or any other institution outside of Utrecht University, some of the guidelines in the Data Privacy Handbook may not be in line with your institutional guidelines. Please consult with your local privacy staff whenever you are in doubt about this.

1.1.1 License and Citation

The Data Privacy Handbook is licensed under a Creative Commons Attribution 4.0 International License. You can view the license here.

When using (parts of) the Data Privacy Handbook in your work, please cite us using the citation provided in the GitHub repository.

1.1.2 Contributions

The Data Privacy Handbook is a collaborative effort, made possible by a large number of contributors. All contributors can be found in our GitHub repository.

Would you like to contribute to this Handbook yourself? Please follow our Contributing guidelines and the Style guide.

1.2 How to use this Handbook

The Data Privacy Handbook aims to make knowledge and solutions on handling personal data *Findable, Accessible, Interoperable, and Reusable* (FAIR) and present them in a practical format.

The Handbook need not be read like a textbook. You are invited to navigate to the topic you need based on the table of contents, or use the guide below.

1.2.1 What are you looking for?

I want to…:

Learn about the GDPR in the context of scientific research

Introduction to the GDPR

Definitions

Plan a GDPR-compliant research project

Designing your research project

Choosing a legal basis

Assessing the risks in your project, for example using a privacy scan, Data Protection Impact Assessment, or Data classification

Informing participants

Obtaining consent

Collaborating on personal data

Setting up agreements

Work safely with personal data

Storing personal data

Finding suitable tools and services

De-identifying personal data

Securely analysing personal data

Sharing personal data during research

Using other approaches to protect personal data, such as encryption, statistical approaches to de-identification, synthetic data, or data donation

Share personal data with others

Sharing data legally

Sharing personal data during research

De-identifying personal data

Securely analysing personal data

Using GDPR-compliant tools and services

Sharing personal data for reuse

Sharing personal data case by case

Learn from other projects

Minimising personal data in a survey

Pseudonymising different types of data

Publishing metadata only

Reusing education data for research purposes

Get help or information

Getting help at Utrecht University

Definitions

References

1.3 Disclaimer

The content presented in the Data Privacy Handbook has been carefully curated by Research Data Management Support, in collaboration with privacy officers and data experts of Utrecht University.

The Data Privacy Handbook is a ‘living’ book that is continually being written, updated and reviewed. Its contents can therefore change, or become outdated or redundant. Hence, the information presented is provided “as is”, **without guarantees of accuracy or completeness**.

As scientific research may differ depending on the discipline, topic, and context, measures needed or taken to ensure GDPR-compliance will vary across research projects. The authors can therefore **not be held responsible, nor accountable** for any negative consequences arising from interpretation and use of the content of the Data Privacy Handbook.

The Handbook does not necessarily constitute a mandatory directive. **For the most up-to-date and official/ authoritative information, please refer to the university website and intranet, to which this Handbook is a hands-on, practical supplement**. Moreover, before implementing the guidance laid out in this Handbook, always seek the advice of your privacy officer or RDM Support to confirm the suitability of any proposed solution to your project.

Throughout the Data Privacy Handbook, links to external webpages may be provided for additional information or assistance. The authors of the Data Privacy Handbook are **not responsible for the content of any such linked webpages**, nor is the content of external webpages necessarily endorsed by Utrecht University.

Utrecht University is committed to sharing knowledge in line with the principles of open science and therefore welcomes readers from outside of the organisation. However, the contents of the Data Privacy Handbook may not be in line with readers' institutions' policies or views. For more authoritative information, these readers should refer to resources from their own institutions.

Chapter 2

Research scenarios

This chapter will outline typical privacy issues and design solutions for several types of scientific research.

These scenarios are as yet a Work In Progress. For now, please consult the rest of the Data Privacy Handbook, or contact your privacy officer if you cannot find the answer here.

Chapter 3

Privacy FAQs

Date of last review: 2023-01-27

On this page you can find Frequently Asked Questions (FAQs) about handling personal data in research. Click a question you have to read its answer.

3.0.1 General questions

When should I be dealing with privacy in my project?

You should think about privacy:

as soon as you are processing personal data. Processing means anything you do with personal data, e.g., collecting, analysing, sharing, storing, etc. The definition of personal data is explained in the chapter What are personal data?.

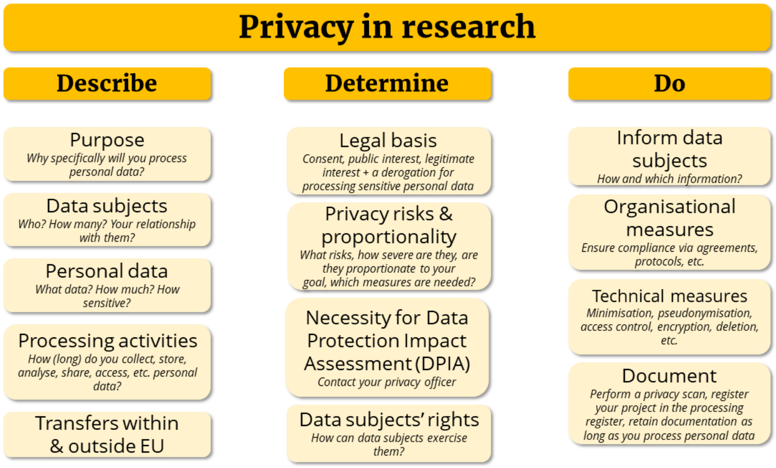
during the earliest stages of your project. This principle is called "privacy by design". It is easier and more effective to address any privacy issues at the design phase of your project rather than having to change your plans later on due to privacy concerns.

When are data truly anonymous?

You can read all about this in the chapters What are personal data? and Pseudonymisation and anonymisation.

What should I consider when handling personal data?

It is best to conduct a privacy scan to check if you work with personal data. The below figure summarises what you need to describe, determine and do as part of



Privacy in research

Describe

Purpose
Why specifically will you process personal data?

Data subjects
Who? How many? Your relationship with them?

Personal data
What data? How much? How sensitive?

Processing activities
How (long) do you collect, store, analyse, share, access, etc. personal data?

Transfers within & outside EU

Determine

Legal basis
Consent, public interest, legitimate interest + a derogation for processing sensitive personal data

Privacy risks & proportionality
What risks, how severe are they, are they proportionate to your goal, which measures are needed?

Necessity for Data Protection Impact Assessment (DPIA)
Contact your privacy officer

Data subjects' rights
How can data subjects exercise them?

Do

Inform data subjects
How and which information?

Organisational measures
Ensure compliance via agreements, protocols, etc.

Technical measures
Minimisation, pseudonymisation, access control, encryption, deletion, etc.

Document
Perform a privacy scan, register your project in the processing register, retain documentation as long as you process personal data

a privacy scan:

My data were collected prior to the GDPR, what rules do I need to follow?

The GDPR applies to all personal data, including those collected prior to the GDPR (May 2018). Therefore, there is really no difference between how personal data should be handled before or after the advent of the GDPR.

My data were collected outside of the EU, does the GDPR apply to them?

Yes, as long as personal data are being processed, and the data controller, data processor, or data subject reside(s) in the European Economic Area, the GDPR applies.

How sensitive are my data?

Personal data can differ in sensitivity, depending on the type of data (e.g., sensitive personal data), of whom the data were collected (e.g., healthy adults, children, patients, elderly, etc.) and on which scale. Data classification and a Data Protection Impact Assessment are useful tools to assess how sensitive the data are.

3.0.2 Procedures and responsibilities

Who is responsible for correctly handling personal data?

Legally, the controller of the personal data is responsible, i.e., the people or organisation responsible for the project activities. If you are an employee at Utrecht University (UU), the UU is legally the controller. The UU however delegates this responsibility to the appropriate employee who is actually in charge of determining why and how personal data are handled. In a research context, this is usually the researcher on the project (e.g., PhD candidate, principal investigator).

What does the procedure look like for researchers at Utrecht University?

All researchers at UU have to write a Data Management Plan. Besides that, many faculties require that a privacy scan is done and ethical approval is obtained. Preferably, a Data Management Plan and privacy scan (which has to sometimes be extended to a Data Protection Impact Assessment) are done (and preferably marked as positive by the relevant data steward/privacy officer) before the ethical review takes place. Once accepted by the ethical committee, you can then start your research project.

How long will the planning process of my research take?

This differs per faculty, but you should count at least 1 month, if not more, to complete all planning activities. In terms of administrative work, you need to reserve time for:

writing a Data Management Plan and having it reviewed (a few days)

filling out the privacy scan and consulting with the privacy officer (a few days). If a DPIA needs to be conducted, this will take more time because the Data Protection Officer also needs to be consulted.

creating information for data subjects and potentially a consent form.

going through ethical review: it can take up to 1 month before a first decision is taken by some faculty review boards, or longer for the Medical-Ethical Review Board.

in some projects, setting up an agreement.

In general, designing your research with correctly processing personal data in mind will cost you less effort In the long run: Start as early as possible!

Doesn't the ethical committee also look at privacy?

Partly, although this differs per UU faculty. In most faculties, there is a collaboration between privacy and ethics. For example, at the Faculties of Social and Behavioural Sciences, the Humanities, and Geosciences, privacy is included in the ethical application, but the privacy aspect of it is outsourced to the faculty privacy officer. For you as a researcher, it is wise to first complete a draft privacy scan, and consult with the faculty privacy officer and only then do the ethical application, so you have already thought about the privacy aspect before the ethical review process starts.

3.0.3 Informed consent

When is parental consent needed?

The GDPR dictates that at least one legal guardian provide consent if you process personal data from children under 16 years old. Note that for medical data and in some faculties, there can be additional requirements, such as obtaining written consent from both parents, and also from the child themselves if the child is between 12 and 15 years old.

Can consent be digital?

Yes, as long as you can demonstrate that consent was obtained, it is valid according to the GDPR. Consent can for example consist of participants ticking a checkbox in a survey tool after reading or watching information about the research. The checkbox should be empty at the start of the survey and not already come "pre-ticked": consent must be actively given. Note that for medical data, consent may have to be provided in writing: always check with your Ethical Review Board.

Where can I find a template consent form?

You can use the minimal list of requirements for an information letter and read through the guidance on informed consent and information to data subjects. Please note that some Ethical Review Boards have specific templates that you should use.

How to balance being complete vs. being intelligible in the information to participants?

The GDPR does not require you to provide all details in the same way to participants. For example, it allows you to layer information, and it requires that you always provide the information in a format that is intelligible for your target audience, which does not necessarily have to be in text. Please refer to the section on Information to data subjects for more information on this. Please note that some Ethical Review Boards have specific requirements on how information should be provided to participants, which have to do with ethical and legal aspects other than the GDPR.

Where, how and for how long should I store my consent forms?

Consent forms have to be stored securely (access-controlled) and separately from the research data, for as long as the research data contain personal data. This can be in digital or physical (e.g., paper) form. Once the personal data are deleted or fully anonymised, the consent forms should be deleted as well. An empty consent form can be stored for longer, for example to check the phrases about re-use. Read more about this in the Data storage chapter.

A participant wants to withdraw their consent. Can I continue to use their data afterwards?

No, once a participant has withdrawn consent, you are obliged to remove any of their data that is under your control and cease any further use of their data from that point onwards. Any processing that occurred *prior* to withdrawal is nevertheless still legal. For example, if the data were published and made publicly available prior to their withdrawal, you are not obliged to take down the entire dataset and seek all individuals that may have downloaded the data subject's data. Another example is if you already analysed the data (but have not yet published the results). In that case, the data have to be deleted, but you do not necessarily need to re-do the analysis. The only important thing is that the data then no longer support the analysis, so for research integrity reasons, you may want to re-do the analysis anyway.Additionally, if you cannot find the participant's data in your dataset because they are deidentified too much, then you are exempt from removing them, unless participants can provide you with information to enable their re-identification.

3.0.4 Legal questions

What if I cannot formulate a specific research question in advance?

It is not always possible in research to be very specific about what the personal data will be used for in the future. In some cases, you can therefore use the concept of "broad consent", where you continuously inform data subjects and enable them to exercise their rights. This is described in more detail here.

I will move to another institution, can I take my research data that contains personal data with me?

Moving data to another institution constitutes a new way of "processing" data and implies that there will be a new (additional) controller of the personal data. This means that you need to take some additional steps, such as ensuring that data subjects are informed about the move, the purpose of the transfer is compatible with the original purpose(s) for data use, both institutions sign an agreement on data protection, use and ownership of the data, etc. What is possible depends largely on the context of your research and the type of data you have: contact your faculty privacy officer for assistance.

When do I have to perform a Data Protection Impact Assessment?

If there is a possibility for a high risk of damages to data subjects, a DPIA is mandatory. This can for example be the case when you observe people in public spaces or process sensitive personal data on a large scale. Note that correctly performing a DPIA can take some time. Contact your faculty privacy officer if you suspect that you may need a DPIA.

Do I need an agreement?

An agreement is usually needed when someone outside of your institution accesses (personal) data that you control. Please refer to the Agreements section to assess whether you need an agreement, and if so, which type.

What is the difference between a Data Transfer Agreement and a Data Processing Agreement?

A data transfer agreement is needed when (personal) data are transferred from one controller to another, and is also recommended to use when data are transferred between departments of a single controller, to delineate the agreed upon responsibilities. For example, in research is it used often when data are shared with other researchers for reuse. A data processing agreement is needed when personal data are transferred from a controller to a processor. For example, it is needed to ensure that an external survey tool protects the university's personal data sufficiently and does not use it for their own purposes, only to provide their survey services. You can read more about these agreements here.

Am I a processor as employee of my university?

No. As an employee you are still determining your own why (research question) and how (methods) of personal data processing. This makes you a controller, acting as an "agent" of the legal controller (your university). Read more on the difference between processors and controllers on the definitions page.

3.0.5 Storing personal data

Where should I store physical personal data?

Physical personal data should be stored in a locked area that only a select group of people has access to. The exact location will depend on the type of data (e.g., consent forms, filled out questionnaires, biomedical samples, etc.), and where you work. If possible, we recommend digitising and then destroying any paper materials in order to have the data in a secure and backed-up location.

Where to store participants' contact information?

Similarly to informed consent forms, you should store contact information on a different location than the research data and well-protected (strict access control, encryption, etc.). For example, store the research data on Yoda, and the contact information in a controlled OneDrive or ResearchDrive folder. Delete the contact information when you do not longer need them (e.g., after the research project has ended).

3.0.6 Sharing, publishing and reusing personal data

Can I publish personal data?

This is not only a privacy issue, but also an ethical one. You can in principle ask consent to publish personal data (either publicly or under restricted access), or in some cases rely on public interest to do so. Because the data will remain protected by the GDPR, anyone (re)using the data will have to abide by the GDPR as well (the requirements travel with the personal data). However, even if you have a legal basis to publish personal data, it still may not always be ethical to do so. For that reason, we recommend always obtaining ethical approval, including when you want to publish personal data. You can read more about sharing and publishing personal data for reuse in the Sharing data for reuse chapter.

How can I share personal data with collaborators?

If the collaborator resides outside of your institute, but within the European Economic Area (EEA) or an "adequate" country, it is possible to share personal data with them, provided that data subjects are informed, there is a (joint controllers) agreement with them, and other safeguards are in place (e.g., pseudonymisation). Please contact your privacy officer if the collaborator is located outside the EEA in a country without an adequate level of data protection.

How can I share data with a third party outside of the EEA?

Personal data can be shared outside of the EEA if one of the following applies:

Participants have given their explicit consent after having been well informed of the risks.

The transfer is necessary for important reasons of public interest.

The data are transferred to a non-EEA country that has been deemed adequate by the European Commission.

The above apply only to "occasional" transfers. For frequent transfers, Standard Contractual Clauses should be drafted, although this requires a greater commitment from the third parties, and may require more in-depth legal assistance to establish.

What should I do if some participants do not consent to sharing their data?

This depends on the identifiability of the data and the legal basis: if it is still possible to identify individuals, then data subjects can withdraw their consent, and you won't be able to share their data for reuse. However, if the data are altered in such a way that you can no longer identify individuals within the dataset, then you can share their data for reuse. Of note, it is not always necessary to ask people their consent for data reuse for scientific purposes - consult your privacy officer. You can read more about this in the Sharing data for reuse chapter.

Can I reuse medical data for research purposes?

You likely can. The GDPR has a derogation that specifies that secondary use for research is "not incompatible with the initial purposes" (art. 5(1)(b)), meaning that it is allowed to reuse data for research, provided that you protect the data sufficiently. As with any research project, we recommend to conduct a privacy scan to assess the legality of your project, and to obtain ethical approval to assess the ethical aspects of your project.

Can I use personal data that are already published by other researchers?

You generally can, depending on the license or terms of use that the dataset has, and assuming that the researcher who published the data had a legal basis to do so. In general, it is possible to reuse personal data for scientific research, as long as appropriate safeguards are in place (art. 89).

Can I reuse contact details for a new study?

This depends on how data subjects were informed about potential reuse of their contact details: can they expect to be contacted again and for this purpose? Note that you should have obtained access to the contact details legitimately too: are you supposed to have access to their contact details in the first place? If you are uncertain about this, ask your privacy officer for help.

3.0.7 Practical questions

I am using hardware to collect personal data. What should I take into account?

There are many security aspects to consider when using hardware (e.g., tablets, cameras, phones, etc.), such as whether and where any personal data is recorded and whether the device is approved by the university, see this link for more information. Make sure that you transfer the data to secure storage as soon as possible and consider measures (such as encryption) that ensure that data are protected if the hardware is lost or stolen. When you use video recording hardware, be mindful of what is recorded, also in the background. For example, be aware when filming around open laptops, documents or vulnerable people.

I want to combine data from multiple sources. How can I do so securely?

There are multiple factors to consider, depending on the type of research, the ownership of the data, involved parties, etc. As a rule of thumb, practice data minimisation, only keep the fields or variables you need. Be mindful of data ownership: if someone else owns the data, keep that dataset separate. For more information and tailored advice, contact RDM Support.

How to generate suitable pseudonyms?

A pseudonym can be a random number, cryptographic hash function, text string, etc. It is important that the pseudonym is not meaningful with respect to the data subjects: a random (unique) number or string is better than a code that contains parts of personal information, because the latter may reveal details about data subjects.

How to pseudonymise qualitative data?

Textual data is often redacted (either manually or using a tool so that identifiable information is removed or replaced with a placeholder text. There are now also tools for masking or blurring video data and distorting audio. Note that sometimes it is not possible to anonymise or pseudonymise qualitative data, because you may lose too much valuable information, or because the data are just too revealing (e.g., face, voice, gestures, posture in video data, language use in audio data). In that case, other measures like access control, safe storage, and encryption may be more suitable.

I am analysing my data in a git repository to ensure reproducibility. How can I make sure I do not accidentally push the data to GitHub?

Before you put your data in your git repository, place a line in the .gitignore file that prevents tracking the data. This way, when pushed to GitHub, the data will not be pushed alongside the other files in the repository - only the folder name will be visible. Please note that if the data were tracked by git before, adding a line to your .gitignore will not prevent the data from being tracked. In this case, it is best to create a new git repository where you add a .gitignore file from the start, and delete all old versions from GitHub if there were any. If you delete the data, add the line to the .gitignore file, and then re-add the dataset, the tracking history from before the .gitignore will still exist and be pushed to GitHub. Sidenote: it is possible to override the .gitignore file by force. This will likely not happen accidentally, but it is important to realise that the .gitignore file is not iron clad. You can read more on the gitignore here.

How to securely send participant data to participants?

In the same protected way as when you would send personal data to fellow researchers. Researchers at Utrecht University can for example use SURF filesender with encryption or share a OneDrive or Research Drive file. Be sure not to share any data from other participants or other researchers!

How to work responsibly with social media data?

See these guidelines (in Dutch) about working with social media data. Every social media platform has different terms and conditions. Read these to see what you are, and are not, allowed to do with the data published on the platform you wish to research.

Where can I find relevant or approved tools?

Researchers at Utrecht University can find tools via https://tools.uu.nl and the intranet. We also curated an overview of several tools to handle personal data in this GitHub repository.

Where can I find privacy-related templates and examples for research?

Please refer to the Documents and agreements chapter or the RDM website. For others, please contact your privacy officer and/or your Ethical Review Board.

3.0.8 Students and student data

Can I reuse educational data (e.g., grades, course evaluations) for my research?

It is possible, but its compliance would have to be documented in a privacy scan to explain why this further processing for scientific purposes is compliant with the GDPR. Please refer to the use case about this topic for an example.

Can I share my research data containing personal data with my students?

Preferably not. Especially in a classroom setting, students should work on anonymised data as much as possible. For thesis students, only share personal data with them as strictly necessary and make sure that the students know how to safely handle the personal data. Additionally, data subjects should be informed that these students will handle their data.

Can I (re)use personal data collected by my students?

You should check what information was given to data subjects to see whether it is possible to reuse the data. In general, if data are deidentified and are going to be used for research, it is possible to make this data reuse legitimate - a privacy scan may be able to demonstrate this.

When students collect personal data, who is responsible for correct handling of those data?

The supervisor is the main person responsible, but students are also co-responsible, especially if they are taking decisions on the data themselves. Students need to comply with their respective obligations and responsibilities to ensure data is kept safe and protected.

Can a student take research data containing personal data with them to publish about them later?

It depends on why this is considered necessary, if data subjects have been informed, if data minimisation and deidentification are applied, etc. If students take data with them, they will probably end up being stored on a free cloud solution such as Google Drive or Dropbox. Make sure your data subjects are informed about this beforehand and realise that obtaining consent will be more difficult. A privacy scan should document why this is compliant with the GDPR.

I am a student, where can I store my data?

If you are student who will be collecting personal data for research, it is the responsibility of your supervisor or course coordinator to supply you with access to an approved storage solution. Please do not use a personal device or commercial cloud solutions like Dropbox or Google Drive to store research data containing personal data. Any “free” commercial solution will scrape and analyse what you store and thus your data are not safe there.

3.0.9 Finding support

Where can I learn more?

Please see the Seeking help page for more information and contact persons for all your questions about privacy, research data management and security.

Who is the Data Protection Officer (DPO)?

The Data Protection Officer (Dutch: Functionaris Gegevensbescherming, FG) oversees an organisation's compliance to the General Data Protection Regulation (GDPR). In research, the DPO is sometimes involved in a Data Protection Impact Assessment and in some cases in possible data breaches. If you work at Utrecht University, you can read more about the DPO's role here.

I have a potential data breach, what should I do?

If you work or study at Utrecht University, please report this as soon as possible to the Computer Emergency Response Team (CERT).

Knowledge Base

The GDPR

This chapter will present the most important definitions, principles and rights of data subjects outlined in the GDPR and how it applies to your research. Most of the practical advice that we provide in this Handbook will be rooted in and builds on the concepts presented here.

3.0.10 Chapter summary

The GDPR is a EU-wide regulation that controls the processing of personal data. If you process personal data, you should:

- Make sure you have a legal basis to process the data. In research, this is often informed consent.
- Be transparent and fair towards data subjects.
- Be specific in which personal data you process and for what purposes. Limit the amount of data you process to what is necessary, and only store the data for that necessary amount of time.
- Protect the confidentiality of the data by incorporating privacy by design into your project from the start.
- Make sure your data subjects can exercise their data subjects' rights, and they know how to do so.

3.1 What is the GDPR?

On this page: gdpr, when privacy, uavg
Date of last review: 2022-07-11

The General Data Protection Regulation (GDPR, Dutch: *Algemene Verordening Gegevensbescherming* [AVG]) is an EU-wide regulation meant to protect the privacy of individuals within a rapidly growing technological society. The

GDPR facilitates the free movement of personal data within the European Economic Area (EEA). Its data processing principles are meant to ensure a fair balance between competing interests – for example, the right to conduct research vs. the right to protect personal data (Articles 13 and 8, from the Charter of Fundamental right of the EU).

3.1.0.1 The GDPR in a nutshell

All articles and recitals of the GDPR can be found online via https://gdpr-info.eu/. The video below highlights some important aspects of the GDPR:

Click to read the English video transcript

The General Data Protection Regulation (GDPR) regulates what we can and cannot do with personal data such as a person's name, sexual orientation, home address and health. This also applies to personal data used in research and education. The regulation consists of 88 pages. Fortunately, the basics are easy to remember in 3 steps:

First, there must be a clear legal basis for processing personal data. This can include consent, a legal obligation, or public interest.

Second, appropriate technical and organisational measures must be taken while processing personal data to ensure maximum privacy.

Lastly, the persons whose data you have collected must always have the option of inspecting, changing, or removing their personal data.

That is the GDPR in a nutshell.

3.1.0.2 When does the GDPR apply?

The GDPR has been applicable from May 2018 onward and applies when:

- you are processing personal data (material scope, art 2).
- the controller or processor of the data *resides* in the EEA (territorial scope, art. 3). This is independent of whether the actual processing takes place in the EEA. In some cases, the GDPR also applies when the controller or processor is not established in the EEA, but is processing data from EU citizens.

If you are collecting or using data that originated from individuals (or is related to individuals), it is very likely that the GDPR applies to your project. You can read more in the chapter What are personal data?.

3.1.0.3 Implementation

While the GDPR is a regulation for the entire EEA, each EEA country can additionally implement further restrictions and guidelines in national implementation laws. The Dutch implementation law is called “Uitvoeringswet AVG (UAVG)”. The UAVG determines, for example, that it is forbidden to process Citizen Service Numbers (BSN), unless it is for purposes determined by a law or a General Administrative Order (AMvB).

3.2 Definitions in the GDPR

On this page: glossary, sensitive data, personal data, process, controller, processor, participant, data subject, special categories, legal ground, legal basis, anonymised, pseudonymised
Date of last review: 2023-07-11

Below, you will find a selection of important terms in the GDPR that you should become familiar with when working with personal data (also included in the Glossary). Click on a term to see the definition.

Data subject

A living individual who can be identified directly or indirectly through personal data. In a research setting, this would be the individual whose personal data is being processed (see below for the definition of processing).

Personal data

Any information related to an identified or identifiable (living) natural person. This can include identifiers (name, identification number, location data, online identifier or a combination of identifiers) or factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of the person. Moreover, IP addresses, opinions, tweets, answers to questionnaires, etc. may also be personal data, either by itself or through a combination of one another.

Of note: as soon as you collect data related to a person that is identifiable, you are processing personal data. Additionally, pseudonymised data is still considered personal data. Read more in What are personal data?.

Special categories of personal data

Any information pertaining to the data subject which reveals any of the below categories:

racial or ethnic origin

political opinions

religious or philosophical beliefs

trade union membership

genetic and biometric data when meant to uniquely identify someone

physical or mental health conditions

an individual's sex life or sexual orientation

The processing of these categories of data is **prohibited**, unless one of the exceptions of article 9 applies. For example, an exception applies when:

the data subject has provided explicit consent to process these data for a specific purpose,

the data subject has made the data publicly available themselves,

processing is necessary for scientific research purposes and obtaining consent is impossible or would require an unreasonable amount of effort.

Contact your privacy officer if you wish to process special categories of personal data.

Processing

Any operation performed on personal data. This includes collection, storage, organisation, alteration, analysis, transcription, sharing, publishing, deletion, etc.

Controller

The natural or legal entity that, alone or with others, determines or has an influence on **why** and **how** personal data are processed. On an organisational level, Utrecht University (UU) is the controller of personal data collected by UU researchers and will be held responsible in case of GDPR infringement. On a practical level, however, researchers (e.g., Principal Investigators) often determine why and how data are processed, and are thus fulfilling the role of controller themselves.

Note that it is possible to be a controller without having access to personal data, for example if you assign an external company to execute research for which you determined which data they should collect, among which data subjects, how, and for what purpose.

Processor

A natural or legal entity that processes personal data on behalf of the controller. For example, when using a cloud transcription service, you often need to send personal data (e.g., an audio recording) to the transcription service for the purpose of your research, which is then fulfilling the role of processor. Other examples of processors are mailhouses used to send emails to data subjects, or Trusted Third Parties who hold the keyfile to link pseudonyms to personal data. When using such a third party, you must have a data processing agreement in place.

Legal basis

Any processing of personal data should have a valid legal basis. Without it, you are now allowed to process personal data at all. The GDPR provides 6 legal bases: consent, public interest, legitimate interest, legal obligation, performance of a contract, and vital interest. Consent and public interest are most often used in a research context.

Anonymous data

Any data where an individual is irreversibly de-identified, both directly (e.g., through names and email addresses) and indirectly. The latter means that you cannot identify someone:

by combining variables or datasets (e.g., a combination of date of birth, gender and birthplace, or the combination of a dataset with its name-number key)

via inference, i.e., when you can deduce who the data are about (e.g., when "profession" is Dutch prime minister, it is clear who the data is about)

by singling out a single subject, such as through unique data points, e.g., someone who is 210 cm tall is relatively easy to identify)

Anonymous data are no longer personal data and thus not subject to GDPR compliance. In practice, anonymous data may be difficult to attain and care must be given that the data legitimately cannot be traced to an individual in any way. The document Opinion 05/2014 on Anonymisation Techniques explains the criteria that must be met for data to be considered anonymous.

Pseudonymous data

Personal data that cannot lead to identification *without additional information*, such as a key file linking pseudonyms to names. This additional information should be kept separately and securely and makes for de-identification that is reversible. Data are sometimes pseudonymised by replacing direct identifiers (e.g., names) with a participant code (e.g., number). However, this may not always suffice, as sometimes it is still possible to identify participants indirectly (e.g., through linkage, inference or singling out). Importantly, pseudonymous data are still personal data and therefore must be handled in accordance with the GDPR.

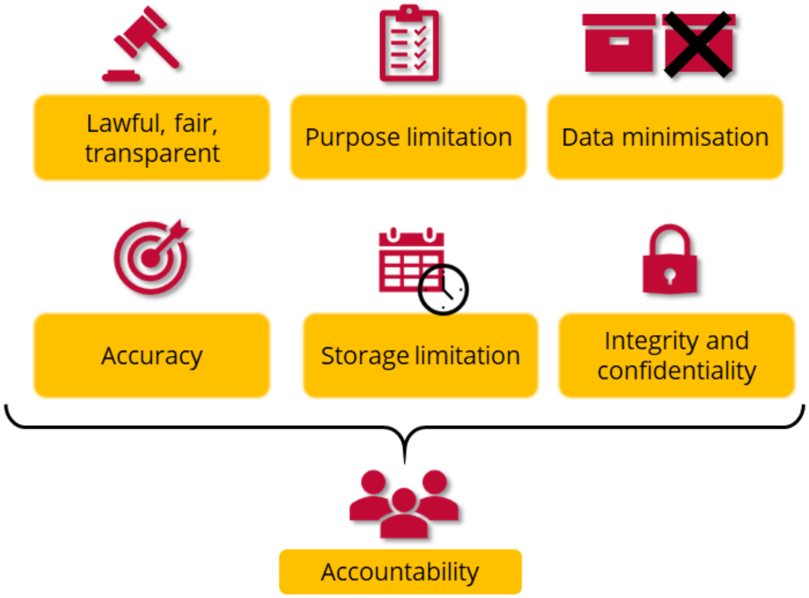
3.3 Principles in the GDPR

On this page: legal basis, legal ground, fair, transparent, purpose, goal, aim, minimise, accurate, storing, storage, safeguards, measures, responsible, responsibility
Date of last review: 2023-07-11

The GDPR has a number of principles at its core which dictate the (method of) data processing. Every type of processing has to comply

with these principles. Understanding these principles is the first step to determining what type of personal data can be collected and how they can processed.

The GDPR principles are explained further below the image. The Design chapter describes how to implement these principles in your research. You can also always contact your privacy officer.



3.3.0.1 1. Lawful, fair and transparent

When working with personal data, your processing should be:

Lawful

Make sure all your processing activities (e.g., data collection, storage, analysis, sharing) have a legal basis. Ideally, you should have determined your processing purposes (e.g., research questions) in advance.

Fair

Consider the broad effects of your processing on the rights and dignity of the data subject.

Give data subjects the possibility to exercise their rights.

Avoid deception in the communication with data subjects: processing of personal data should be in line with what they can expect.

The processing of personal data should not have a disproportionate negative, unlawful, discriminating or misleading effect on data subjects.

Transparent

Be transparent in the communication to your data subjects about who is processing the personal data (controllers, processors), which personal data are processed, as well as why and for how long, and how data subjects can exercise their rights. The information provided should be unambiguous, concise, easily accessible and relevant and shared with data subjects before the start of your research.

3.3.0.2 2. Purpose limitation

You can only process (i.e., collect, analyse, store, share, etc.) personal data for a specific purpose and only for as long as necessary to complete that purpose. For example, if you communicated to data subjects that you would use their personal data only to answer your specific research question, you cannot further share the personal data for new research questions, as these would be additional processing purposes. This means that you need to **plan what you will do with the (collected) personal data in advance and stick to that plan in order to be GDPR-compliant**.

3.3.0.3 3. Data minimisation

You can only process the personal data you need to for your predefined purpose(s), and not more just because they may “come in handy later”. This principle makes sure that, for example, in the event of a data breach, the amount of data exposed is kept to a minimum.

3.3.0.4 4. Accuracy

The accuracy of personal data is integral to data protection. Inaccurate data can be a risk for data subjects, for example when they lead to a wrong treatment in a medical trial. You therefore need to take every reasonable step to remove or rectify data that is inaccurate or incomplete. Moreover, data subjects have the right to request that inaccurate or incomplete data be removed or rectified within 30 days.

3.3.0.5 5. Storage limitation

You can only store personal data for as long as is necessary to achieve your (research) purpose. Afterwards, they need to be removed. If the personal data are part of your research data (and not, for example, to simply contact data subjects), you are allowed to store (archive) them for a longer period of time, provided necessary safeguards are in place. This is an exemption that applies to data storage for scientific archiving purposes. You need to inform the data subjects on this storage duration beforehand.

If identification of the data subject is no longer needed for your (research) purposes, you do not need to keep storing the personal data just to comply with the GDPR, even if it means your data subjects cannot exercise their rights (art. 11).

3.3.0.6 6. Integrity and confidentiality

You have to process personal data securely and protect against unauthorised processing or access, loss or damage. To this end, you should put in place apropriate organisational and technical measures.

3.3.0.7 7. Accountability

The controller is ultimately responsible for demonstrating GDPR-compliance. As a researcher working with personal data, you are representing your institution (e.g., Utrecht University) and you should therefore be able to demonstrate that you process personal data in a compliant manner. Additionally, you should also have some knowledge of data protection so that you can implement the right measures into your research project.

3.4 Legal bases for working with personal data

On this page: legal basis, legal ground, consent, public interest, legitimate interest, secondary use
Date of last review: 2023-07-11

You can only process personal data if you have a **legal basis** to do so, which should be registered, among other information, in the processing register and communicated to data subjects. There are 6 possible legal bases which are outlined below. In research, the legal bases ‘informed consent’, ‘public interest’ and to some extent ‘legitimate interests of the controller’ are most often used.

For different purposes in your research project, a different legal basis may apply. For example, you may contact data subjects before they start participating based on a legitimate interest and use informed consent for collecting, storing, analysing and publishing the data.

3.4.1 Legal bases suitable for research

Informed consent

Informed consent is the most frequently used legal basis in research and is often not only a legal (GDPR-consent), but also an ethical obligation (e.g., METC informed consent). When using informed consent, you should be able to demonstrate that the data subject was informed and has given consent, and for which purpose(s) they gave their consent. In all cases, consent has to be freely given, specific, informed and unambiguous. Please refer to the Informed consent section for guidance on applying informed consent in your research.

Public interest

Public interest is sometimes used in research when the research is shown to clearly benefit the public good or fulfills a public task. In essence, public interest can be used for research that is conducted by employees of public institutions, when their research interest has been recognised by an official authority. For example, conducting research at Dutch universities has been officially recognised in the Higher Education and Scientific Research Act to be a public task. Public interest is often used when consent is not a good option. For example, it may be impossible or impractical to obtain consent when performing public observations or social media research. Or when participants actually do not have a free choice, such as in clinical trials when participants would experience significant disadvantages when not participating.

If you want to use public interest as a legal basis, you need to assess the necessity and proportionality of your processing. Additionally, you need to demonstrate that the interests of data subjects do not override your research interests. To do so, please contact your privacy officer to assess whether you can use this legal basis in your research.

Legitimate interest of the controller

Legitimate interest is often used by companies to process personal data necessary for the functioning of their own company, e.g., processing user data for fraud prevention, or keeping a registration system to provide better services. In research, legitimate interest is often used for processing activities that have no direct research purpose. For example, this can be the case when you need to collect contact information to approach data subjects to participate, and you can only obtain their consent for participating in your research after contacting them. Since contacting data subjects is a prerequisite to perform your research, it can be in the university's legitimate (research) interest to process their contact information.

To evaluate whether you can use legitimate interest as a legal basis, you always need to weigh the interests of the controller (e.g., Utrecht University) and the data subjects in a Legitimate interest assessment. Please contact your privacy officer to assess whether you can use this legal basis in your research.

3.4.2 Legal bases not suitable for reseach

Processing is necessary because of a legal obligation of the controller

This basis is not suitable for research. As an example, Utrecht University has to share tax data with the Dutch tax administration in order to comply with tax legislation.

Processing is necessary for the performance of a contract

This basis is not suitable for research. As an example, Utrecht University has contracts with its employees, which require it to manage the employees' financial data.

Processing is necessary to protect a person's vital interests

This basis is generally not suitable for research. If processing someone's personal data is crucial to their health or even life, that processing is allowed under the GDPR.

3.4.3 Further processing for research purposes

It may happen that you want to process personal data for other purposes than previously specified (e.g., because you formulated an additional research question), or you want to reuse previously collected personal data in your research. In these cases, it may be possible to make use of article 5(1)(b), which states that "further processing for [...] scientific purposes shall [...] not be considered to be incompatible with the initial purposes". Basically, this means that you can reuse personal data, that were previously collected for other purposes, for scientific research purposes. This is only allowed if you put in place sufficient safeguards to protect the personal data, inform data subjects, and allow them to exercise their rights (art. 89). "Further processing" is not strictly a legal basis. Instead, it functions as a way to legitimise *further* processing of personal data (which was previously collected for a different purpose, using one of the six legal bases) for research purposes.

Public interest, legitimate interest, and relying on further processing are ways to meet your *legal* requirements for processing personal data, but not necessarily your *ethical* requirements: you may still need consent if demanded so from an ethical perspective. Before you rely on any of these, you should first assess whether they are indeed suitable with your faculty privacy officer, and determine whether your research interests outweigh the privacy rights of the data subjects.

3.5 Data Subjects' Rights

On this page: rights of participants, right, withdrawing consent, delete data
Date of last review: 2023-07-11

The GDPR provides data subjects with several rights that gives them a relatively high degree of control over their own personal data. Below, we list these rights and how you can apply them in your research:

Right to be informed

Data subjects need to be clearly informed about what you are doing with their personal data (a.o. art. 12). This usually happens via an information letter. This right does not apply if your research will be seriously harmed by meeting it and if you haven't obtained the personal data directly from the data subjects themselves.

Right of access

Data subjects have the right to access a copy of the personal data you have on them and to know what you are doing with that personal data and why (art. 15).

Right to rectification

Data subjects have the right to correct and complement the personal data that you have of them (art. 16).

Right to erasure/Right to be forgotten

Data subjects have the right to have their personal data removed (i.e., equivalent to the right to withdraw consent, art. 17). This right does **not** need to be granted if:

the personal data are published and need to be archived for validation purposes.

it would seriously obstruct the research purpose(s).

it would hinder complying with a legal obligation or carrying out a task in the public interest.

If the personal data have already been made public or shared, you need to take reasonable measures to inform other users of the data of the erasure request. A privacy officer can help you with this.

Right to restriction of processing

Data subjects have the right to have you process less of their personal data (art. 18), for example if their personal data are inaccurate or your processing of it is unlawful or no longer needed.

Right to data portability

Data subjects have the right to have their personal data transferred to another party in a "structured, commonly used and machine-readable format" (art. 20).

Right to object

Data subjects have the right to object to what you are doing with their personal data. This right applies when the processing is based on legitimate or public interest (art. 21). In case of objection, you have to stop your processing activities and thus delete any data you have from the particular data subject, unless you can demonstrate concrete grounds for overriding the data subject's rights (e.g., excluding the data subject would substantially bias your results).

In Dutch scientific research, it is possible to exclude the rights of access (art. 15), rectification (art. 16), and restriction of processing (art. 18), so that data subjects cannot exercise those rights (UAVG art. 44). If you wish to do this, please first consult with your privacy officer.

3.5.0.1 How can data subjects exercise their rights?

Data subjects need to be **informed** about their rights and who to contact in order to exercise them, including when you use a legal basis other than informed consent. In research, this is usually done via a privacy notice or information letter, which states a contact person responsible for handling questions and requests.

Incoming requests need to be **coordinated with a privacy officer**, so that they can be picked up in accordance with the GDPR. Additionally, at Utrecht University, data subjects can always contact privacy@uu.nl (Legal Affairs) for requests or complaints.

3.5.0.2 What to do when receiving a request concerning data subjects' rights?

You have to provide a substantive response to the data subject **within 30 days**, in the same way as you received the request. Depending on the complexity and number of requests, the response period may be extended by 2 months. In that case, you must inform the data subject about this extension (including the motivation) within one month. If needed, you can (and sometimes should) ask for additional information to confirm the data subject's identity.

For granting requests about data subjects' rights, there should be a procedure in place, in which you should at least consider:

- how you are going to retrieve the data (e.g., using a name-number key)
- who is responsible for granting the request and informing the data subject about it (e.g., a data manager)
- how the request is going to be granted, for example how they will be sent securely (access, portability), removed (forgotten, object, restriction) or corrected (rectification)

For larger projects, it may be wise to put a Standard Operating Procedure (SOP) in place.

3.5.0.3 What if the data have already been anonymised?

The principles of data minimisation and storage limitation are considered more important than keeping personal data just for the sake of identification (art. 11). Therefore, when receiving a request about anonymised data, you can make it clear that you cannot retrieve the data subject's personal data, because they have been anonymised. In this case, **the data subject cannot exercise their rights anymore**. If you can still retrieve the data subject's personal data in some way (i.e., when data are pseudonymised), you are **obliged to retrieve them**. In order to do so, you can (and sometimes should) ask for additional information that can confirm the data subject's identity.

Chapter 4

What are personal data?

In order to know whether you should comply with the GDPR in your research project, the first question to answer is: do you process personal data? To answer this question, we need to know: (1) What exactly are personal data, and (2) how do you know if you are working with personal data in your research?

4.0.1 Definition of personal data

According to the GDPR, personal data are “any information relating to an identified or identifiable natural person” (art. 4(1)):

- **Natural person**: Data by themselves (numbers, text, pictures, audio, etc.) are not inherently personal. They only become personal when they refer to or relate to a **living individual**. When data refer to an organisation, deceased person, or group of individuals, they are not considered personal data under the GDPR.
- Data are personal if they **relate** to an individual. This means practically anything that someone is, has said or done, owns, may think, etc.
- The person should be **identified or identifiable**. This is the case not only through **directly** identifying information, such as names and contact information, but also through **indirectly** identifying information, for example if you can single someone out or identify them by combining datasets (see the next page).

4.1 How to assess whether data contain personal data?

On this page: sensitive data, privacy-sensitive, personal data, when is data privacy-sensitive, identifiability, identifier
Date of last review: 2022-08-23

Whether your data contain personal data depends on which data you are collecting (nature) and under which circumstances (context). A date like "12 December 1980", is not personal data – it is just a date. However, that date becomes personal data if it refers to someone's birthday.

In assessing whether data are personal, you should take into account all the means that you and others may **reasonably likely** use to identify your data subjects, such as the required money, time, or (future) developments in technology (rec. 26).

Data can be identifiable when:

- They contain **directly identifying information**.

  For example: name, image, video recording, audio recording, patient number, IP address, email address, phone number, location data, social media data.

- It is possible to **single out** an individual

  This can happen when there are unique data points or unique behavioural patterns which can only apply to one person.

  Examples:

  You have a data subject who is 2.10 meters tall. If this is a unique value in your dataset, this distinguishes this person from others and thus can make them identifiable.

  You have a data subject who only follows far-right accounts on Twitter. If they are the only one in your dataset who do so, this distinguishes this person from others and can make them identifiable.

- It is possible to **infer information** about an individual based on information in your dataset

  For example:

  Inferring a medical condition based on registered medications.

  Guessing that someone lives in a certain neighbourhood based on where they go to school.

- It is possible to **link records** relating to an individual.

  This can happen when combining multiple variables within your dataset (e.g., demographic information, indirect identifiers). However, it can also happen when combining your dataset with other datasets (the “Mosaic effect”). In that case, your data still contain personal data, even if the data in your own dataset are not identifiable by themselves.

Linkage is often possible with demographic information (age, gender, country of origin, education, workplace information, etc.) and indirect identifiers (pseudonyms, device ID, etc.), for example:

In the year 2000, 87% of the United States population was found to be identifiable using a combination of their ZIP code, gender and date of birth. You can see for yourself on this website.

An agricultural company’s Uniek Bedrijfsnummer (UBN) can be used to search for the address of the company in the I&R mobile app. Often, this address is also the owner’s home address.

Geographical data tracking individuals are particularly sensitive because of the multiplicity of data points. This video nicely explains why.

- De-identification is still **reversible**.

  This often happens when data are pseudonymised, but there is still a way to link the pseudonymised data with identifiable data, for example when a name-pseudonym key still exists.

You can assume that you are processing personal data when you collect data **directly** from people, even if the results of that collection are anonymous. But also when you use data that are **observed or derived from people**, even if those data were previously collected, made public or used for non-research purposes.

In short, even if you cannot find out someone’s real identity (name, address), the data you process can still contain personal data under the GDPR. Besides the examples mentioned here, there are many other examples of personal data. If you need help assessing whether or not your data contain personal data, please contact your privacy officer.

4.2 Special types of personal data

On this page: sensitive personal data, sensitive data, special category, special categories, politics, race, ethnicity, religion, philosophy, dna, genetics, genes, fingerprint, physical condition, mental illness, sexual identity, gender identity
Date of last review: 2022-08-23

There are a few special types of personal data that are worth taking note of: special categories of personal data, and otherwise sensitive personal data. These types of personal data have additional requirements. If you want to process them, please contact your privacy officer first.

4.2.1 Special categories of personal data

The GDPR explicitly defines seven 'special categories of personal data'. It is information that reveals:

- racial or ethnic origin
- political opinions
- religious or philosophical beliefs
- trade union membership
- genetic or biometric data when meant to uniquely identify someone
- physical or mental health conditions
- sex life or sexual orientation

It is in principle **prohibited** to process these types of personal data, unless an exception applies (art. 9). For example, it is allowed to process these if:

- Data subjects have provided explicit consent to process these data for a specific purpose.
- Data subjects have made the data publicly available themselves
- Processing is necessary for scientific research purposes (incl. historical and statistical purposes) and it is impossible or would take an unreasonable amount of effort to obtain explicit consent (UAVG art. 24).

Even if you can make use of one of these exemptions, special categories of personal data warrant additional security measures to make sure they are protected. Always contact your privacy officer if you intend on processing these types of data.

The Dutch Code of Conduct for Health Research (p.68) specifies a number of exceptions for health researchers in which explicit consent for processing special categories of personal data may not be necessary.

4.2.2 Data that are otherwise sensitive

Other types of data can also be sensitive, because they can carry higher risks for the data subjects. These types of data can either not be processed at all, or only under certain circumstances. Either way, they require additional security measures. Always contact your privacy officer if you intend on using these types of data.

Examples are:

- Financial data
- Data about relationship problems

- Data that can be misused for identity fraud, such as the Dutch Citizen Service Number (BSN). In principle, the BSN cannot be used in research at all.
- Criminal or justice-related data: they can only be processed under governmental supervision or when a derogation exists in national legislation (art. 10).

Chapter 5

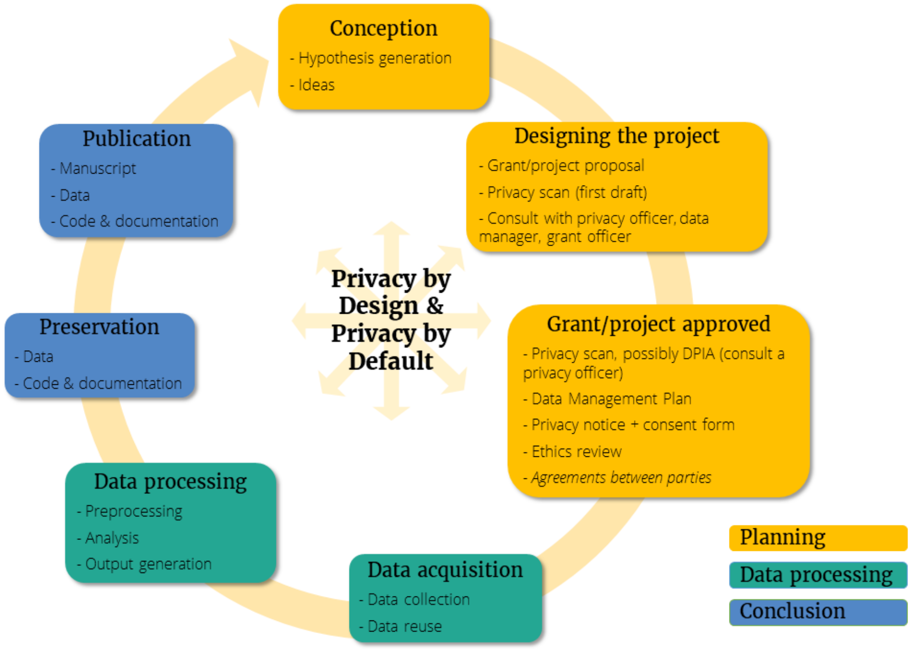
Designing your project

On this page: privacy by design, start early, preparation
Date of last review: 2022-10-31

Research projects typically go through a number of stages: conception, proposal, planning, execution, publishing, preservation, etc. If you work with personal data, you should think about how you will protect those data throughout all those stages. To do so, the concepts of Privacy by Design and Privacy by Default (art. 25) are important:

- **Privacy by Design** in research means that your project integrates personal data protection right from the beginning, all the way throughout the project, and even afterwards. It should not be an afterthought: Privacy by Design is a key feature of the project, permeating all phases of a research project.
- **Privacy by Default** in research means that any questions, tools, or methods you use in your research should process as little personal data as necessary by default, and that you share the personal data only with those who really need access.

To get proper support in designing your project, it is important to contact your privacy officer early on, preferably already in the conception or design phase. The privacy officer will help you go through the different stages smoothly, and eventually save you time and effort. They can help you review and possibly adjust your plans, determine the appropriate protection measures, and determine whether you need to perform a more elaborate Data Protection Impact Assessment.



5.1 Privacy by Design strategies

On this page: safeguards, measures, technical, organisational, procedure, design, access control, minimisation, transparency, pseudonymisation, abstraction, information, accountability, rights
Date of last review: 2022-10-31

To incorporate the concepts of Privacy by Design and Privacy by Default into your project, the approach of **privacy design strategies** (Hoepman, 2022) offers a way to make the GDPR principles more concrete. Hoepman distinguishes 8 strategies that you can apply to protect the personal data in your research: minimise, separate, abstract, hide, inform, control, enforce, and demonstrate. Below, we explain what these mean and how you can apply them.

The GDPR does not prescribe *which* specific measures you should apply in your project, only that they should protect the personal data *effectively*. Which measures will be effective, will depend on your specific project, the risks for data subjects, and the current progress in technology (i.e. will the data be protected on the long haul?).

Data-oriented strategies

Minimise
Separate

Abstract
Hide

Process-oriented strategies

Inform
Control
Enforce
Demonstrate

5.1.0.1 Minimise

Limit as much as possible the processing of personal data, for example by:

- Collecting as little data as possible to reach your research purpose.
- Collecting only personal data from the amount of individuals necessary.
- Preferably not using tools that automatically collect unnecessary personal data. If possible, prevent tools you do use from doing so (Privacy by Default). For example, the survey tool Qualtrics can automatically register location data, which can be turned off by using the "Anonymize Responses" option.
- Removing personal data when you no longer need them. Remove them from repositories, data collection tools, sent emails, back-ups, etc. (see also the Storage chapter). Use directly identifying information only if you legitimately need them, for example to keep in touch with data subjects or to answer your research question.
- Pseudonymising or anonymising personal data as early as possible.
- Use portable storage media only temporarily.

Back to top

5.1.0.2 Separate

Separate the processing of different types of personal data as much as possible, for example by:

- Storing directly identifying personal data (e.g., contact information) separately from the research data. Use identification keys to link both datasets, and store these keys also separately from the research data.
- Separating access to different types of personal data. For example, separate who has access to contact information vs. to the research data.
- Applying secure computation techniques, where the data remain at a central location and do not have to be moved for the analysis.

Back to top

5.1.0.3 Abstract

Limit as much and as early on as possible the detail in which personal data are processed, for example by:

- Pseudonymising or anonymising the data.
- Adding noise to the data, e.g., voice alteration in audio data.
- Summarising the data to simply describe general trends instead of individual data points.
- Synthesising the data, e.g., for sharing trends in the data without revealing individual data points.

Back to top

5.1.0.4 Hide

Protect personal data, or make them unlinkable or unobservable. Make sure they do not become public or known. You can for example do so using a combination of:

- Using encryption, hashing or strong passwords to protect data. Consider using a password manager to avoid losing access to the data.
- Using secure internet connections and encrypted transport protocols (such as TLS, SFTP, HTTPS, etc.). Do not connect to public WiFi on devices containing personal data.
- Applying privacy models like Differential privacy, where noise is added to individual data points to hide their true identity.
- Only providing access to people who really need it, and only for the necessary amount of time and with the necessary authorisations (e.g., read vs. write access; only the relevant selection of personal data, etc.). Remove authorisations when access is no longer required.
- Encrypting and regularly backing up data on portable storage media.
- Keeping a clear desk policy: lock your screen and store paper behind lock and key when you leave your desk.

Back to top

5.1.0.5 Inform

Inform data subjects about the processing of their personal data in a timely and adequate manner, for example by:

- Providing information via an information letter or privacy notice on a project website.

- Providing verbal explanation before an interview.
- Obtaining explicit consent via an informed consent procedure.

Back to top

5.1.1 Control

Give data subjects adequate control over the processing of their personal data, for example by:

- Specifying a procedure and responsible person in case data subjects want to exercise their data subject rights.
- Providing data subjects with a contact point (e.g., email address) for questions and exercising their data subject rights.

Back to top

5.1.2 Enforce

Commit to processing personal data in a privacy-friendly way, and adequately enforce this, for example by:

- Using only Utrecht University-approved tools to collect, store, analyse and share personal data.
- Entering into agreements with third parties if they are working with UU-controlled personal data. Such agreements will make sure everyone will treat the data up to UU-standards.
- Always keeping your software up-to-date and using a virus scanner on your devices.
- Appointing someone responsible for regulating access to the data.
- Always reporting (suspicions of) data breaches. At UU, contact the Computer Emergency Response Team.
- If needed, drawing up a privacy and/or security policy that specify roles and responsibilities and best practices on how personal data are handled throughout a project.
- Using a Trusted Third Party when linking individual data from different sources together.

Back to top

5.1.3 Demonstrate

Demonstrate you are processing personal data in a privacy-friendly way, for example by:

- Registering your research project in the UU processing register (once available).
- Performing a Privacy Scan and storing it alongside the personal data.
- Performing a Data Protection Impact Assessment (DPIA) for projects that have a high privacy risk for the data subjects.
- Keeping information for data subjects and (signed) informed consent forms on file. This is not needed if you can fully anonymise the data: then you should delete the (signed) consent forms as well.

Back to top

Chapter 6

Risk Assessment

When you work with personal data, you need to make sure that you correctly collect, store, analyse, share, etc. those data to avoid harm to data subjects. To do so, it is important to gain insight in:

- **The risks involved**:
  Security risks occur when data are unexpectedly less available, less correct, or there is an unintended breach of confidentiality. They need to be mitigated by implementing integrity and confidentiality into your project. *Privacy risks* exist when your use of (personal) data, either expectedly or unexpectedly, affects the interests, rights and freedoms of data subjects. These can be Data Subjects' Rights under the GDPR, but also other fundamental rights, such as the right to equality and non-discrimination, the right to life and physical integrity, freedom of expression and information, and religious freedom. In practice, we consider it a privacy risk if your processing of personal data can result in physical, material, or non-material harm to data subjects. Privacy risks should be mitigated by implementing **all** data protection principles into your project. When the risks for data subjects are high, an in-depth risk assessment in the form of a Data Protection Impact Assessment is needed.

- **The data classification**: a classification of the data (low, basic, sensitive, critical) that is based on the risks for data subjects and the damages to an institute or project when data are incorrectly handled, there is unauthorised access, or data are leaked. This classification affects the security measures you need to take (e.g., which storage solution you choose, whether you need to encrypt the data, etc.).

Based on the risks you identified and the classification of the data, you can then implement safeguards to mitigate the risks.

Privacy risks can occur in any stage of your research project (see also Solove, 2006). If the image does not show correctly, view it online.

6.1 How to assess privacy risks?

On this page: risk, security, assessment, harm, damage, dpia, threat, secure, measure, safeguard, protect, plan, probability, likelihood, impact
Date of last review: 2023-04-18

Before you start your research project, it is important to consider the risks and their severity for data subjects in your project. This assessment will inform you on which (additional) safeguards to put in place to mitigate the risks.

Privacy and security risks are usually outlined in a privacy scan or Data Protection Impact Assessment, and purely security risks in a data classification. If you create an algorithm that can affect people, an "Impact Assessment Fundamental Rights and Algorithms" may be required or combined with any of the before mentioned assessments.

6.1.1 Risk assessment step by step

When going through the below steps, take into account at least the following risk scenarios:

- **Data breach** (unintended security risks): someone unauthorised gains (or keeps) access to personal data, or personal data are lost due to a security incident.
- **Inability for data subjects to exercise their rights**: for example, data subjects have not been (well-)informed about data processing, there is no contact person to ask for data removal, or there is no procedure in place to find, correct or remove data subjects' data.
- **Intrusion of personal space**: for example, you observe data subjects in a place or at a time where/when they would expect a sense of privacy (e.g., dressing rooms or at home). If there is secret or excessive observation, people may feel violated and stifled.
- **Inappropriate outcomes**: the outcomes of your research project may also impact data subjects, for example when it induces discrimination, inappropriate bias, (physical or mental) health effects, but also when a lack of participation denies data subjects beneficial treatment effects.

1. Outline which and how much (personal) data you use, how, and for what purposes

   This is usually one of the first steps of a privacy scan.

2. Is there a project with similar data, purposes, methods and techniques?

   If there are projects that are the same or very similar to your project, you can reuse relevant work from their privacy scan, or if applicable, Data Protection Impact Assessment (DPIA). Naturally, you should adjust sections that do not apply in your own project. If you're not sure of any existing projects similar to yours, ask your privacy officer or colleagues.

3. List possible harm to data subjects and others

   Make an overview of the possible harm that could occur to data subjects and others if any of the risk scenarios occurs. These could be:

   Physical harm Damage to someone's physical integrity, such as when they receive the wrong medical treatment, end up as a victim of a violent crime, or develop mental health problems such a depression or anxiety.

   Material harm Destruction or property or economic damage, such as financial loss, career disadvantages, reduced state benefits, identity theft, extortion, unjustified fines, costs for legal advice after a data breach, etc.

   Non-material harm

   Social disadvantage, for example damage to someone's reputation, humiliation, social discrimination, etc.

   Damage to privacy, for example a lack of control over their own data or the feeling of being spied on. This can happen when you collect a lot of personal data, or for a longer period of time (e.g., with surveillance, web applications).

   Chilling effects: when someone stops or avoids doing something they otherwise would, because they fear negative consequences or feel uncomfortable.

   Interference with rights: using personal data may violate other fundamental rights, such as the right to non-discrimination or freedom of expression.

4. Estimate the risk level without safeguards

   After listing the possible harm, you should determine the risk level of each harm occurring. The risk level depends on:

   the **impact** of the harm: what is the effect of each of the 4 scenarios above on the data subject and others (major, substantial, manageable, minor)?

   the **likelihood** of the harm occurring: this depends on the circumstances of your project, such as: what and who can cause the harm to occur? How easily are mistakes made (e.g., how easily will an unauthorised person gain access)?

   It is important to first determine the risk level in case you do not implement any safeguards. This will be your risk level if all those safeguards fail. The higher this initial risk, the more you should do to mitigate it.

5. Determine the safeguards you can use to mitigate the risks

   In many cases, it is possible to mitigate the risks by implementing organisational and technical measures. The higher the risks, the more and/or stricter measures should be in place to mitigate them. You can find some relevant measures in the Privacy by Design chapter, and on the example page in this chapter.

6. Determine the residual risk after implementing safeguards

   By implementing safeguards, you are decreasing the likelihood of the risks occurring. If the risk is still unacceptably high, even after implementing safeguards, you should:

   Modify your processing to reduce the impact of potential damages (for example, refrain from collecting specific data types), or

   Implement more or better measures, reducing the likelihood of any harm occurring.

It will always be difficult to quantify risks. Therefore, it is largely the argumentation that can provide context in how the risk level was determined. The same harm may in one project be very unlikely to occur, while in another it may be very likely: **context matters**!

6.2 What are high-risk operations?

On this page: high-risk, large risk, dpia, assessment, mandatory
Date of last review: 2023-04-18

The GDPR requires a Data Protection Impact Assessment (DPIA) to be conducted when the risks in your project are high, considering “the nature, scope, context and purposes” of your project (art. 35(1)). More practically, you need to do a DPIA when two or more of the criteria from the European Data Protection Board apply to your project, or – if the processing occurs in the Netherlands - when one or more of the criteria from the Dutch Data Protection Authority (English UU translation) applies to your project.

6.2.1 Examples of high-risk scenarios

You systematically use automated decision making in your project (art. 35(3))

For example:

- You use an algorithm to analyse health records and predict patients’ risk of complications.

- You use an algorithm to analyse students' test scores and learning patterns, to make personalised recommendations for coursework or additional resources.
- You use an algorithm to detect fraudulent activity.

You process special categories of personal data or criminal offense data on a large scale (art. 35(3))

For example:

- You amplify bodily materials into pluripotent stem cells, cell lines, germ cells or embryos (see the Dutch Code of Conduct for health research, 2022).
- You analyse social media data to study political opinions and religious beliefs.
- You investigate criminal records from all currently incarcerated individuals (note that such a project is likely subject to additional restrictions).

You publicly monitor people on a large scale (art. 35(3))

For example:

- You use traffic data and GPS devices to monitor people's behaviour in traffic.
- You use CCTV footage to study public safety.

You collect a lot of personal data, or from a large group of people (EDPB, 2017)

For example:

- You collect data on psychosocial development in twins annually for over a decade.
- You collect genomic data to study the genetic basis of a specific disease.
- You keep a database with contact information from thousands of people.

You use new techniques or methods for which the effects on data subjects or others are not yet known (EDPB, 2017)

For example:

- Machine learning algorithms.
- Internet of Things.
- Virtual or Augmented Reality.
- Natural Language Processing.
- Human-computer interaction.

Your research involves groups that are vulnerable or touches a vulnerable topic (EDPB, 2017)

For example:

- You perform video interviews with children talking about abuse.
- You interview refugees about their home country.
- You perform in-depth interviews with employees about their job satisfaction.
- You perform a diary study among mentally ill patients.
- You collect data from homosexual individuals in a country where homosexuality is forbidden or can lead to discrimination.
- You perform research among a population with (severe) distrust towards scientific research(ers) or who have difficulty understanding your research.

There is a high chance of incidental findings in your research (Dutch Code of Conduct for health research, 2022)

For example:

- You collect neuroimaging data from patients who likely have a brain tumour.
- You investigate genetic data from vulnerable subjects that indicates a risk for disease.

When you suspect that you may need a DPIA, or when you are not certain whether your project needs one, please contact your privacy officer.

6.3 Data classification

On this page: BIV classificatie, CIA triad, data classification, information security, IT system
Date of last review: 2023-04-18

In order to determine which IT solutions are suitable for processing personal data (e.g., storage or analysis platforms), a classification of your data is needed. That data classification can then be paired to the classification given to IT solutions. Institutes will determine for which data classification certain IT solutions are suitable. For example, at Utrecht University (UU), the classification levels are: low, basic, sensitive or critical. If your data are classified as “critical”, you are not allowed to use an IT solution that is only suitable for “sensitive” data.

To classify data, you determine how important it is to keep the data Confidential, correct (Integrity), and Available. Below you can find some guidance on determining the risk level for each of these. Note that this guidance is based on the UU data classification, but your institute may adhere to a different form of the classification.

Data classification can be done for all types of data, not only personal data. Personal data would simply score “higher” on the Confidentiality aspect.

6.3.1 Classification levels

Confidentiality

How confidential are the data?

- Low:
  - Anonymous data, or data that are already publicly available, from less than 50 people.
  - Direct colleagues.
  - No third parties and software involved.
  - No reputation loss when data are lost.
- Basic:
  - Non-public basic personal data such as name, (email)address, etc.
  - Personal data obtained directly from data subjects.
  - Personal data from a moderate number of data subjects (> 50 - 200).
  - Sensitive personal data from a small number of individuals.
  - Third parties are involved but they are located inside the EEA.
- Sensitive:
  - A data leak would lead to reputation damage to you and the university.
  - You are bound to patents or contractual agreements.
  - Sensitive personal data from a moderate number of data subjects (e.g., personality data, financial data).
  - Non-sensitive personal data from a large number of data subjects (> 10.000).
  - Personal data enriched with external resources.
  - Far-reaching process automation.
  - Non-targeted monitoring.
  - Relatively new technology.
- Critical:
  - Any project that carries high risks for data subjects or others:
    - Highly sensitive personal data (e.g., biometric identification data, genetic data).
    - Personal data from a very large number of data subjects (> 50,000).

        - Vulnerable subjects (e.g., minors, disabled, undocumented, persecuted groups).
        - Processing happens (partly) outside of the EEA without an adequacy decision.
    - Life-threatening research.
    - There are far-reaching contractual obligations.
    - A data leak would lead to exclusion from future grants.

Integrity

How important is it that the data are correct and can only be modified by authorised individuals?

- Low: Incorrect data would be an inconvenience and/or require some rework.
- Basic: Incorrect data would invalidate research and/or require significant rework.
- Sensitive: Incorrect data would invalidate multiple research projects, could cause reputational damage to you and the university, or lead to significant contractual violations.
- Critical: Incorrect data could have far-reaching contractual obligations, exclusion from future grants or life-threatening research.

Availability

How important is it that the data are available? When would it be a problem; if the data are not available for an hour, a day, a week…?

- Low: Losing (access to) the data would be inconvenient and/or lead to rework.
- Basic: Losing (access to) the data would invalidate research and/or require significant rework. Not having access to the data would cause significant delays and could incur costs up to 250.000 EUR.
- Sensitive: Losing (access to) the data would terminate or hugely delay multiple research projects, could cause significant reputational damage to you and the university, lead to significant contractual violations or individuals not being able to access their sensitive personal data.
- Critical: Inaccessible data could have far-reaching contractual obligations, cause damages in excess of 1.500.000 EUR, including exclusion from future grants or losing/not being able to access potentially life-threatening data.

Please note that a classification may be lower or higher than indicated in the examples, depending on your specific context. Please contact your privacy officer to help you classify your data. You can also contact Information Security for questions about data classification and security measures.

6.4 Examples of risks and how to mitigate them

On this page: risk example, safeguards, organisational and technical measures, protection, protective, security, data breach
Date of last review: 2023-04-18

Below you can find a list of common privacy and security risks in research and how you can mitigate them:

- Unwarranted access to personal data
- Loss of personal data
- Unintended collection of personal data
- Invalid legal basis
- Risks for data subjects

6.4.1 Unwarranted access to personal data

Someone tries to gain access to personal data

Use storage and analysis systems that are suitable for your data classification, e.g., systems that are managed by your institute and/or encrypted.

Apply protection strategies described here.

A previous team member still has access (e.g., a copy on their personal device, a working account)

Enforce a protocol in which team members who leave need to remove all their copies of the data and are denied access to the data and shared folders (on- and offboarding). Periodically review and update all users/rights. Make someone responsible for this process.

A team member shares the data with a third party

Put in place a protocol or non-disclosure agreement that makes team members aware that this is not allowed, or make sure that a data transfer agreement is in place.

Make sure that team members do not have access to data that they do not need access to.

A password is leaked

Use systems that apply multifactor authentication.

Change your password regularly or immediately when it is compromised, and have your team members do the same.

Back to top

6.4.2 Loss of personal data

A device is lost or defective (e.g., laptop, USB stick)

Protect the device with a password.

Encrypt the device or the data on it.

Delete unnecessary copies of the data on the device as soon as you've made a back-up on a more stable and secure system, such as university-managed storage facilities.

Enable removing data from the device from a distance.

Paper data are lost

Avoid collecting data on paper altogether, or only collect the necessary information.

Store the paper data in a central and access-controlled location, scan the documents as soon as possible, store the scans on a backed-up storage medium and destroy the paper records (securely).

The dataset is deleted accidentally

Use a storage system that has back-up functionality, or if not available, make regular manual back-ups of the data.

A system error causes temporary loss of or access to data

If you are not using centrally managed IT solutions, regularly check if back-ups are being done as expected and have protocols in place on how to restore back-ups.

If the time-out takes a significant amount of time, discuss with your privacy officer whether you need to inform data subjects about it: they cannot exercise their rights during that time.

The organisation is hit by a ransomware attack

Enforce a security protocol that emphasises secure data practices, such as:

Do not download data from unknown sources.

Be careful when installing software, preferably only install software from the institutional software catalogue.

Create awareness of what phishing looks like and to report phishing immediately to the Computer Emergency Response Team.

Back to top

6.4.3 Unintended collection of personal data

Data subjects give more, or more sensitive information about themselves than intended/needed

Offer data subjects the possibility to review what information they provided.

Offer the possibility to withdraw consent in a later stage.

Use a data collection protocol to prevent this from taking place.

Remove the unnecessary information from your dataset.

Data subjects give (sensitive) information about others

Use a data collection protocol to prevent this from taking place.

Offer data subjects the possibility to review what information they provided.

Remove the unnecessary information from your dataset.

Consider the risks for those others vs. your own research benefits: if the interests for the other people are more important, you should delete or anonymise the information.

Personal data are collected unintendedly

This can happen when a survey tool automatically collects additional data such as IP addresses. You can sometimes turn this off, and otherwise must remove the data as soon as possible after collection.

Back to top

6.4.4 Invalid legal basis

Data subjects were not informed in a way that is understandable for them

This can be a risk with vulnerable subjects, such as children or psychiatric patients but also with data subjects from different cultures. Make sure the information to data subjects is easy to understand, consider other forms than text (e.g., orally). You could even test this with a sub-group of data subjects. Moreover, we recommend going through an ethical review to consider these aspects more in-depth.

Data subjects could not be (fully) informed because it would harm your research project

If fully informing data subjects can negatively affect your research project (art. 14(5)), we recommend going through ethical review and extensively debriefing data subjects after your project, including a possibility to withdraw consent or to object to the processing. In case of secretive research (heimelijk onderzoek), please contact your privacy officer: this requires an in-depth privacy scan.

Data subjects do not know that their data are used for research

This can happen for example in web scraping or archival research. In principle, you need to inform the data subjects directly. If this takes an unreasonable amount of effort, place a link to a privacy statement on a place that those data subjects likely visit (e.g., social media). Point at a possibility to object to your processing.

Consent cannot be demonstrated

Use a system that registers the consent (e.g., a survey tool, an interview recording), preferably with the date of providing consent. If your research involves a survey, make sure data subjects cannot enter the survey itself if they have not ticked the "consent" box(es). Store the consent declarations for as long as you retain the personal data. Do so securely, but separated from the research data.

Data subjects do not want to sign a consent form

Consider whether you actually need a signature. If you do not use real names or a pseudonym unconnected to real names, using a signature would lead to the unnecessary processing of personal data, and a checkbox will likely suffice.

Contact your privacy officer to consider using public interest as a legal basis instead of consent. Note that data subjects still need to be informed properly.

If you have to use consent, consider the format of consent: for some groups oral consent may work better than written consent.

Consent may not be freely given because you do research in your own organisation

Consider whether you can rely on public interest instead of consent: contact your privacy officer for assistance.

If you need to use consent, try to distance yourself from the data subjects. For example, if your data subjects are students, have someone other than the teacher perform the data collection and/or analysis, or investigate a department other than your own, and prevent the management of the department of interest from getting involved in your project.

Back to top

6.4.5 Risks for data subjects

Your research has a stigmatising effect on the data subjects due to incorrect, unclear or opaque selection criteria

Describe clearly how the data subjects are selected.

Due to a small sample size, data subjects are easily identifiable

If you cannot increase the sample size, put in place protection measures to protect the identity of the data subjects.

Data subjects put themselves in harm's way by participating

Balance the interests of the data subjects vs. those of your research project and go through ethical review.

Collect the data in a physically safe location.

Put in place protection measures like anonymisation, minimisation, blurring, etc. to hide and protect the identity of the data subjects.

Clearly inform data subjects what their participation entails and obtain their explicit consent.

If applicable, inform local authorities and obtain formal permission to perform your research.

Back to top

Chapter 7

Documents & Assessments

This chapter addresses privacy-related documentation that is most commonly used within research. For each type of documentation, it explains what it entails, when to use it, and how.

The following documents are addressed:

- Information to data subjects
- Informed consent forms
- Privacy scan
- Data Protection Impact Assessment
- Legitimate interest assessment
- Data Transfer Impact Assessment
- The processing register
- Several agreements, such as the non-disclosure agreement, data processing agreement, data transfer agreement, joint controller agreement, and data use agreement.

7.1 Information to data subjects

On this page: informed consent, informing, information, transparency, transparent, privacy notice, information letter, privacy policy
Date of last review: 2022-10-07

A privacy notice is any information given to data subjects about what is happening with their personal data. In research, a privacy notice is usually combined with general information about the research project and often with an informed consent form, to satisfy both privacy and ethical concerns. Generally, the aim of a privacy notice is to inform

data subjects on how and why their data are being processed. Providing that information is the "cornerstone of data subjects' rights", as without it, data subjects cannot exercise their other privacy rights.

7.1.1 When to use a privacy notice?

Informing data subjects is always required, for **all legal bases** (so not only when you use informed consent). Being properly informed is a data subject's right in itself (art. 12): it is necessary so that data subjects can exercise their other rights (e.g., right to be forgotten, right to object, etc.).

You need to inform data subjects **before** you start collecting or otherwise processing their personal data (so before the start of your research project). If you share personal data with an external party, you should inform data subjects at the latest when first sharing those data with that external party.

When you use personal data from another source, you have to inform your data subjects within a month after obtaining their data (art. 14), except if:

- they have already been properly informed elsewhere
- this would involve a disproportionate effort
- this would seriously impair your processing purposes (e.g., if you cannot answer your research question anymore, art. 14(5)).

7.1.2 Content and examples of privacy notices

Below you can find a list of items to include in your information to data subjects (Template) and some example sentences to (not) include in your privacy notice (click to expand).

Template information letter UU Template information letter for WMO research

Example sentences

Bad promise

Alternative

"After the project ends, we will delete all of your data, so that you will not be identifiable anymore."

"After the end of the study, we will delete the code linking your data to your name. We will store your de-identified data for 10 years for integrity purposes."

"Your data will be fully anonymised before they are shared with others."

"We will remove personal information that could reasonably identify you before we share any files with other researchers."

"All data that you will provide will be kept strictly confidential and will not be shared further."

"The main researcher will keep a link that identifies you to your coded information. They will keep this link secure and available only to the selected members of the research team."

"Your data will only be accessible by the research team, and no one else."

"We will only share your de-identified data with other researchers if they agree to treat your data confidentially and only after approval from the original research team."

"You can withdraw your consent from this study at any time up until the end of the research project. If you withdraw your consent, we will delete all your data from our dataset immediately."

"You can withdraw your consent from this study at any time, without stating a reason why and without any repercussions. Please inform the researcher about your decision. We will then delete any personal data referring to you that we still have, where this is still possible."

7.1.3 Form of a privacy notice

The format of the privacy notice is also crucial. Even if you include all necessary components in your privacy notice, it will **not be GDPR-compliant** if you fail to provide the information in an appropriate form, shape and time.

The information you provide to data subjects should be:

- **Clear and understandable** A privacy notice is not a legal document, so do not write it like you would write a legal contract. The information should be understandable for data subjects and it should have a clear and concrete meaning. For example, avoid using words like "may", "some", "often" and write active and short sentences. Tip: try these writing tips, this language tester, or use this simplified information sheet (all in Dutch).
- **Easily accessible** Data subjects should be able to find the information easily. For example, publish the privacy notice on your project website, give participants a copy, or provide a QR-code or short URL. Even if you cannot inform data subjects directly, you should make an effort to inform them and put the information somewhere they will likely come across (such as a website or on social media).
- **Via multiple channels (when appropriate)** Textual information sheets are by no means the only way to inform data subjects. If appropriate, you can provide the information via other channels too, e.g., oral statements, images (example), audio, video (example), etc. For

some data subjects, such other channels of informing can lead to a better understanding of your processing activities.

- **Layered (when appropriate)** To balance being complete with being understandable, you can layer the information you provide. For example, provide concise information up front and provide more detailed information elsewhere (e.g., via a link or dropdown menu).

If you are uncertain about the level of intelligibility and transparency of the information, you can test these, for example through user panels, readability testing, or by interactions with data subjects themselves (or their representatives).

7.2 Informed consent

On this page: consent, consent form, informed consent form, legal basis
Date of last review: 2022-11-15

Of the 6 possible legal bases to process personal data, informed consent is currently the one most often used in research. With the term consent, we mean the process of data subjects deciding whether or not to agree to specific statements, such as a statement to participate in a research project.

7.2.1 Consent step-by-step

1. Determine if consent is the legal basis you need

   Determine if consent is the legal basis you need for your research: there are other legal bases besides consent which can sometimes be more suitable in a research context.

   In some situations, consent is likely the only way to process data, for example, if you want to process special categories of personal data, or if you process personal data from people who are incapable of giving consent or from children under 16 years old. In the latter case, the GDPR requires to obtain additional obtain consent from a legal representative (e.g., parent), and there are additional requirements when your research falls under the Dutch Medical Research Involving Human Subjects Act.

2. Consider if you meet all requirements for consent

   If you need to use consent as a legal basis, consider if you meet all requirements listed below. If you do not, consent is not a valid legal basis, and you should consider another one.

3. Determine what you will ask consent for

   Determine what specifically you are asking consent for. If you cannot determine a specific purpose, for instance because your research question is not yet entirely clear, contact your privacy officer to consider obtaining broad consent.

4. Prepare information for data subjects

   Prepare a privacy notice or information letter for data subjects to inform them before asking for their consent.

5. Obtain demonstrable consent

   Different forms of consent are valid. Note that often a signature is not required.

6. Keep the consent forms available

   Treat the consent declarations as personal data: store them separately and securely from the research data, and for as long as your research data contain personal data.

Note that the term “consent” is used both in the GDPR as well as in an ethical context. As a legal basis, data subjects give consent to process their personal data (e.g., “I consent to my data a, b, c be used for purpose x, y, z”). In an ethical context, consent is a safeguard to give data subjects more control over their personal data, and makes sure they participate voluntarily in the research project (e.g., “I have read the information and agree to participate under the conditions described”). Thus, it can happen that consent is not be the best legal basis to use, but should still be used as an ethical requirement.

7.2.2 Requirements for valid consent

Under the GDPR, consent is only valid when it is **all** of the below (art. 4, art. 7, rec. 32, rec. 42, rec. 43; click to expand):

Freely given

Data subjects should have an actual voluntary choice and should not experience negative consequences if they don’t consent or withdraw their consent. Moreover, they should not be pressured to provide consent, and so there cannot be a power imbalance between the controller (e.g., researcher) and data subjects (rec. 43). Some examples:

Consent is not a valid legal basis when the researcher is also a teacher and asks their students to participate, who depend on the teacher for a good grade.

Consent in a clinical trial is not a valid legal basis when patients are asked to participate in the trial, but the choice to participate affects their treatment plan or treatment outcome.

Consent can still be used for children and persons legally incapable to provide consent when their legal representative(s) provide the consent.

Specific

Data subjects should know as specifically as possible what they are asked to consent to. Separate processing purposes therefore require explicitly separate consent (rec. 32, rec. 43), and accompanying specific information that will allow the data subjects to decide if they consent or not. If consents for multiple purposes are necessary for your research, you can combine those. Some examples:

Combined consent may be possible to collect, store, analyse, and share personal data with your collaborators – all actions are needed to answer your research question.

Separate consent is needed for conducting a survey vs. for conducting a subsequent interview, if participation in that interview is not required for your research project.

Separate consent is needed for the current research project vs. for contacting data subjects for future research projects.

Separate consent is needed to use personal data to answer a research question vs. to link different sources of data together to do so (Gedragscode Gezondheidsonderzoek, 2022.

Separate consent is recommended to make the personal data available for reuse (describe the conditions under which this will be allowed).

Informed

Data subjects need to be clearly and accessibly informed about which personal data are processed and why, and about their rights (see Information to data subjects). Data subjects should be able to access this information easily (also after they have provided consent).

Unambiguous and affirmative

It should be clear what data subjects are providing consent for, using a clear, affirmative statement. Importantly, “silence, pre-ticked boxes or inactivity” do not constitute valid consent (rec. 32): consent should be active.

Retractable

Data subjects have the right to withdraw their consent, meaning their personal data cannot be used for the research purpose anymore and have to be removed where possible. Withdrawing consent should be as easy as providing consent. It is important to make the distinction with the right to stop participating at any time (usually an ethical obligation), because the latter implies that the data collected up until that point can still be used for the research project.

7.2.3 What forms of consent are valid?

The way you obtain consent may differ per research project and can depend on how you interact with your data subjects. The only requirement is that it should be demonstrable and registered in a reliable manner. Some examples:

- Ticking a box (**not** pre-ticked!)
- Writing or replying to an email ("I agree to be interviewed")
- Filling in an electronic form
- Audio- or video-recorded consent (separate it from the research data!)
- Signing a paper document (not usually necessary)

7.2.3.1 To sign or not to sign?

Signatures in consent forms are rarely needed. In fact, if you are only processing *pseudonymised* research data, you will only collect unnecessary personal data by obtaining a signature (art. 11), and a checkbox should be sufficient. In order to link the consent form with the data subject, you should include the pseudonym on the consent form (the identifier you will use for the participant, e.g., "part-001"). Inform your participants of this pseudonym; they can use it to exercise their rights under the GDPR, such as for withdrawing their consent.

Only when the identity of the data subjects will be used in the process (e.g., clinical trials), a signature may make sense or be required. For example, if your research is subject to the Dutch Medical Research Involving Human Subjects Act (WMO), different requirements may apply.

7.2.4 Demonstrating (valid) consent

As long as you process personal data, you should be able to demonstrate that the data subjects consented to that processing (rec. 42). So as long as you analyse, use, store, archive, etc. the personal data, the proof of consent needs to be retained. It is preferable to store the proofs separately from the research data. If you collected consent on paper, it is best practice to scan the consent forms and securely delete the paper version after having made sure the scanning went well. Only after there is no personal data anymore (e.g., after fully anonymising the dataset), you can remove the proof of consent.

7.2.5 Broad consent in research

In research, it can sometimes be difficult to formulate very specific research questions in advance. In this case, you may be able to formulate the research purposes on a more general level and obtain consent for these more general purposes (EDPS, 2020; Deutsche Datenschutzkonferenz, 2019). However, you can do this only as long as:

- data subjects can give consent to only part of the research and easily withdraw consent (rec. 33).
- data subjects are kept informed as specifically as possible about what will happen to their personal data. As soon as you know more, you should also inform data subjects in more detail. Your use of the personal data should fall within the line of expectation from data subjects.
- you use additional protection measures, for example:
  - obtain ethical approval for using the data for new research questions.
  - offer a consent withdrawal possibility before using the data for new research questions. This is especially relevant when it is still possible to reliably identify data subjects in the dataset.
  - make sure the data are not transferred to countries outside of the EEA, unless one of the derogations from GDPR Chapter V applies (e.g., adequacy decision, standard contractual clauses, explicit consent for transfer).
  - enforce specific requirements for access the data, e.g., "research in general" is not a sufficiently specific purpose for reuse of the personal data.
- you document your considerations and ask for help from a privacy officer.

Broad consent under the GDPR needs to be distinguished from "General consent" as defined by the Dutch Code of Conduct for health researchers, that is: for medical research, different requirements may (additionally) apply.

7.2.6 Examples and templates

Note that all examples below assume that they are preceded by sufficiently specified information.

Template in Qualtrics Examples CESSDA

Example sentences

Good example sentences:

"I consent to the collection and use of my personal data to answer the research question described in the information letter."

"I consent to linking the new research data to data previously collected about me in this research project."

"I agree that research data gathered for the study may be published or made available provided my name or other identifying information is not used."

"I understand that the research data, without any personal information that could identify me (not linked to me) may be shared with other researchers."

Bad example sentences:

"Any information I give will be used for this research project only and will not be used for any other purpose": this restricts all future uses of the data, including sharing the data with your collaborators, performing analyses for new research questions, and sharing the data for reuse. It's preferred to tell data subjects how their data can be safely used in different ways.

"I do not give consent to share my data": this sentence is ambiguous and may confuse data subjects.

"I acknowledge that the personal data collected by the researcher belongs to the university and that I have no rights in the research performed on it": it is not allowed to deny data subjects their data subjects' rights.

7.3 Privacy scan

On this page: privacy scan, pre-DPIA, DPIA-light, design, data management plan for privacy, risk assessment, planning
Date of last review: 2023-02-14

A privacy scan is an initial risk assessment that helps you delineate how you will handle and protect the personal data in your research project ("a Data Management Plan for personal data", also known as "pre-DPIA", "DPIA-light", or "privacy review"). It contains information on your research question(s), which personal data you process and from which data subjects, how you use the personal data (e.g., will you share them) and which protective measures you apply, your legal basis, how data subjects can exercise their data subjects' rights, and a preliminary assessment of the risks for data subjects.

The purpose of a privacy scan is to:

- Make a preliminary assessment of the risks of your project for data subjects.

- Implement Privacy by Design and Privacy by Default into your project.
- Fulfil the principle of Accountability by documenting your project.
- Identify whether a full Data Protection Impact Assessment (DPIA) is needed.

7.3.1 When to use a privacy scan?

Whenever you use personal data in your project, we recommend to complete a privacy scan in consultation with your privacy officer to make sure your data are well protected throughout your project. As the privacy scan is a planning document, much like a Data Management Plan, it is preferable to fill it out as **early as possible before you start collecting data**, to prevent unforeseen or costly changes to the design of your research project.

- Treat the privacy scan as a living document: update it if anything changes in your design.
- Retain the privacy scan as long as you retain personal data.

Note that there is some overlap in content with some Data Management Plans and research protocols (e.g., that of the CCMO). The privacy aspects in a privacy scan are just more extensive.

7.3.2 Examples and templates

Example template

7.4 Data Protection Impact Assessment

On this page: DPIA, GBEB, risk assessment, high risk, protection measures, safeguards
Date of last review: 2023-02-14

A Data Protection Impact Assessment (DPIA) is an instrument to identify and mitigate privacy risks associated with processing personal data in a project. Whereas a privacy scan is recommended for all projects processing personal data, a DPIA is required by the GDPR if your processing of personal data poses a high risk to the rights and freedoms of data subjects or others (art. 35). This can be the case, for example, when you process personal data from vulnerable groups such as children or patients, sensitive personal data, or a large amount of personal data (see also the risk assessment chapter).

A DPIA is very in-depth, and requires an official advice from the university's Data Protection Officer (DPO). We strongly recommend contacting your privacy officer early on: they can best estimate whether a DPIA is necessary and to identify any approved DPIAs that may be useful for your project to reuse. And they have to be involved in performing the assessment anyway.

7.4.1 The process of performing a DPIA

1. Contact your privacy officer as early as possible to assess the necessity to carry out a DPIA.

2. Work together with your privacy officer, and possible other stakeholders like security officers, to assess your design and risks and complete the DPIA.
3. When finished, the DPIA will be sent to the Data Protection Officer (DPO) for advice. Their considerations will also need to be documented.
4. You may need to adjust your research design and update the DPIA accordingly.
5. In case of a negative DPO advice, you should ask your head of department or faculty dean for permission to go ahead with your project.
6. Regularly update the DPIA when there are changes in your research project.
7. Retain the DPIA for as long as you retain personal data.

7.4.2 Examples and templates

Example template Norea Example template Dutch government

An example case from Utrecht University about social safety in the Dutch House of Representatives is described on the UU intranet (privacy considerations) and on the UU website (news message).

7.5 Legitimate interest assessment

On this page: proportionality, necessity, proportional, necessary, balancing test, legitimate interest, legal basis
Date of last review: 2023-02-14

If you plan to use legitimate interest as a legal basis in your research project, the GDPR requires that you assess the balance between your interests and those of your data subjects (art. 6). In such an assessment, you consider:

- the purpose of your research: what is your interest? The interest (purpose) must be real, concrete and direct. In research, enabling you to perform your research is usually a legitimate purpose.
- whether your processing is:
  - necessary: can you reasonably achieve your goal in a more privacy-friendly way?, and
  - proportionate: how many people will be affected, to what extent and how intrusive is your processing?
- your interests vs. those of data subjects (balancing test), e.g., can data subjects expect you to process their data this way, what is the impact of your processing on data subjects, your project, and society, and which safeguards can you put in place to protect data subjects' interests?

7.5.1 How to do a legitimate interest assessment?

Assessing the legitimacy of your processing is part of the privacy scan or, if applicable, a DPIA. If you do not use a privacy scan or DPIA and/or you have not performed this assessment (yet), but you do rely on legitimate interest, please contact your privacy officer as soon as possible to perform a privacy scan anyways.

Please note: once your interest is assessed as being "legitimate", this is not a free pass to do whatever you want with the data: you still need to incorporate Privacy by design into your project, for example by adequately informing data subjects, protecting the personal data, and allowing data subjects to exercise their rights.

7.5.2 Examples and templates

Template in the absence of a privacy scan/DPIA

7.6 Data Transfer Impact Assessment

On this page: data transfer, third-country transfer, sharing outside EU, EEA, risk assessment
Date of last review: 2023-02-14

A Data Transfer Impact Assessment (DTIA) is a risk analysis that is needed when personal data are transferred to third countries. A DTIA is not an official GDPR document by itself, like the DPIA, but instead is **usually part of, or a supplement to, a DPIA**.

7.6.1 What is a third-country transfer?

In legal terms, a transfer exists when personal data controlled by one party are accessible to another, irrespective of whether the data are physically sent to that party. An international/third-country transfer exists when the party that can potentially gain access is based in a country outside the European Economic Area (EEA) which does not have an adequacy decision from the European Commission.

In GDPR terminology:

- No transfer: a researcher who is connected to an EEA-organisation - and thus has to abide by the GDPR - accesses data that are stored at that EEA-organisation from a "non-adequate" country (provided safeguards are in place to prevent other parties from gaining access).

- Third-country transfer: if personal data are stored at the servers of a non-EEA party, or uses a cloud provider that has servers both in- and outside the EEA.
- No third-country transfer: the party with whom the data are shared is already subject to the GDPR (e.g., the party is situated in Germany or Italy).

7.6.2 Goal and content of a DTA

The goal of a DTIA is to:

- assess the risks of:
  - the data receiver not being able to provide the promised level of protection.
  - local regulations preventing the removal or returning of the personal data after use.
  - local authorities accessing the personal data (il)legitimately.
- determine the appropriate safeguards to protect the data during the transfer.

The DTIA should ideally contain:

1. the context of the data transfer (which data are transferred, how, where?)
2. under which safeguards (art. 46) the data will be transferred (e.g., Standard Contractual Clauses)
3. how effective the safeguards will be (risk analysis)
4. which additional safeguards are needed to ensure a sufficient level of data protection
5. a final decision on whether or not the data can be transferred

As this is a relatively new topic in data protection land, please contact your privacy officer for assistance with a DTIA or for questions about third-country transfers .

7.6.3 Examples and templates

Example questions

7.7 Processing register

On this page: registry, processing activities, project register, administration
Date of last review: 2022-01-05

Any EU-based organisation is required to keep a registry of processing activities from within their organisation (art. 30). If you are performing research with human subjects, the odds are that your research project will have to be registered in such a register as well. Such a register should contain who are processing the personal data, for what purpose, which personal data are processed, with whom the data are shared, and how the data will be protected.

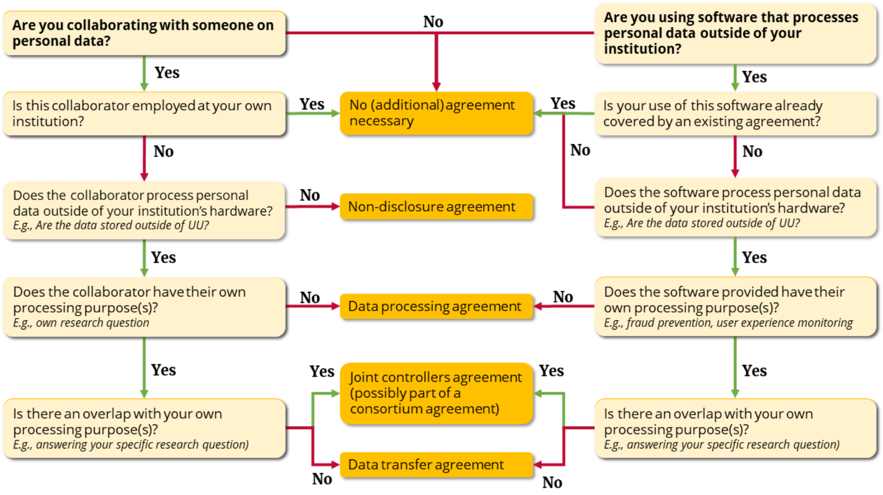
Notably, the GDPR does not prescribe what such a register should look like exactly, and thus every institution has their own way in which the processing register is implemented and managed. Ask your local privacy officer how your institution registers research projects that process personal data.

At Utrecht University, a university-wide processing register is currently being developed. For the time being, other administrative systems and documents such as the Data Management Plan, the privacy scan, the DPIA and/or the ethical application may function as a processing register. Please ask your privacy officer for more information about how your faculty handles this.

7.8 Agreements

On this page: agreement, contract, sharing, transfer, sharing outside EU, EEA, controller, processor, provider, data owner
Date of last review: 2022-12-15

There are many types of agreements that can – and sometimes should – be set up during a research project. Which one you need depends on the purpose of transfer, where the data are transferred to, and the external party's role in the research project. The flowchart below gives an indication which agreement should be used in which situation. Under the figure, you can also find a short explanation about each individual type of agreement.


Are you collaborating with someone on personal data?
No
Are you using software that processes personal data outside of your institution?
Yes
Yes
Is this collaborator employed at your own institution?
Yes
No (additional) agreement necessary
Yes
Is your use of this software already covered by an existing agreement?
No
No
No
Does the collaborator process personal data outside of your institution's hardware?
E.g., Are the data stored outside of UU?
No
Non-disclosure agreement
Does the software process personal data outside of your institution's hardware?
E.g., Are the data stored outside of UU?
Yes
Yes
Does the collaborator have their own processing purpose(s)?
E.g., own research question
No
Data processing agreement
No
Does the software provided have their own processing purpose(s)?
E.g., fraud prevention, user experience monitoring
Yes
Yes
Yes
Joint controllers agreement (possibly part of a consortium agreement)
Yes
Is there an overlap with your own processing purpose(s)?
E.g., answering your specific research question)
Data transfer agreement
Is there an overlap with your own processing purpose(s)?
E.g., answering your specific research question)
No
No

Quick links to: NDA DPA DTA Joint controllers agreement DUA SCCs

7.8.1 How to set up an agreement?

In order to set up an agreement, you should always get in touch with your Research Support Office (RSO) or privacy officer to get the ball rolling!

- Any agreement should be signed by someone who is authorised/mandated to do so. Usually this is a research director or faculty dean, but rarely you yourself. The RSO can tell you who in your case is mandated to sign the agreement.
- An agreement is not a replacement for consent or any other legal basis. It is a safeguard to make sure all parties involved treat the data safely and in accordance with the GDPR.

7.8.2 Non-disclosure agreement

A Non-Disclosure Agreement (NDA), or Confidentiality agreement, is an agreement that makes sure that either the receiver of the data or both parties handle data with care. Often, an NDA is meant to make sure that the receiving party keeps the data they get access to safe and processes the data according to specific guidelines. In research, it is often used between university researchers and students who perform research on their behalf. In this case, it is sometimes necessary to use an NDA, because students are not (always) bound to confidentiality through a contract with the university, whereas the researchers are.

Model NDA Utrecht University (two-sided) Example NDA (one-sided)

7.8.3 Data processing agreement

A data processing agreement (DPA) is mandatory when you transfer personal data to a third party who fulfills the role of a processor (art. 28(3)). In other words: a person or organisation that processes personal data on your behalf, without having a say as to why or how the data are processed. Important components of a DPA are a description of the data that are being shared, why they are being shared, what the third party can or cannot do with them, for how long, and what happens in case of a personal data breach. For example, the third party cannot use the data for their own purposes and is required to keep the data safe and report any potential data breaches to you.

A DPA is most often used when you use an application or tool that processes personal data. Examples of these are survey tools, analysis tools, transcription tools, documentation tools and data repositories. With many parties that offer such tools, there is already a processing agreement in place at the UU-level, and so they are already safe to use, see the Tooladvisor.

Utrecht University template (UU only) SURF template

7.8.4 Data Transfer Agreement

A data transfer agreement (DTA) is advisable when you transfer data to a third party who will (re)use the data for their own purposes, without having an active role in your research project. It is used often when this third party is an external institution, but is also recommended when the third party is someone from within your own institution. A DTA is used to ensure that both parties are aware of their responsibilities and are bound to do what the agreement says. It contains a description of the personal data that are being shared, why they are being shared, under which legal basis, and how the data should be protected by each party.

A DTA can be used when (for example):

- you want to share personal data for reuse purposes with other researchers. This also requires adding statements on the terms of use, although these terms of use can also be separately registered in a Data Use Agreement.
- you are using a software tool, and the software provider wants to run analytics on the personal data you process in their tool. This makes the software provider a controller, with a separate purpose from your own (e.g., answering your research question). Note that you likely already have a Data Processing Agreement with the software provider, and thus a clause in that DPA can also suffice in such cases.

Health-RI template Example Utrecht University

7.8.5 Joint controllers agreement

A joint controllers agreement is mandatory when you work together with another controller on the same personal data, and you have common purposes (why) and means (how) of processing. In a joint controllers agreement, the respective responsibilities of both (all) parties are formalised, e.g., who informs data subjects, who is the contact point for data subjects, and how are data kept secure (art. 26). In research, this happens most often in a consortium, where multiple institutions participate in a research project. Therefore, a joint controllers agreement is often part of the consortium agreement, in which also topics other than the processing of personal data are formalised (e.g., intellectual property rights, how data are shared, etc.).

Examples Erasmus University Rotterdam template SURF template

7.8.6 Data Use Agreement

A data use agreement (DUA), or user agreement, is basically a custom license for your dataset. It specifies the terms and conditions under which the receiver of the data can (re)use the data and is therefore in many ways similar to a Data Transfer Agreement. For example, it may contain statements on what the receiver can do with the data, and how the original data owners should be attributed. It can also state that the receiver must comply with the GDPR and cannot try to reidentify data subjects. DUAs are often used in data repositories (e.g., custom terms of use in DataverseNL) or as part of a Data Transfer Agreement, and often (but not always) when an open license is not suitable (e.g., with personal data).

Example Donders Institute

7.8.7 Standard Contractual Clauses for international transfers

Standard Contractual Clauses (SCCs, art. 46) are model clauses that have been pre-approved" by the European Commission to include in agreements. They are specifically meant as a (sometimes necessary) safeguard when personal data are transferred to processors and controllers outside of the EEA (i.e., third-country transfers). This is because the SCCs contain (among others) a list of minimal necessary safeguards that the receiving party should implement to ensure that the personal data are properly protected. SCCs only have to be used in specific situations, and they should ideally be preceded by a Data Transfer Impact Assessment that identifies SCCs as an appropriate measure. Please contact your privacy officer if you have questions about them.

SCCs for international transfers

There are also SCCs for transfers to processors *within* the EEA (art. 28(7)). These can be included in a Data Processing Agreement. Because they are standardised clauses, including these can make it easier to finalise a DPA.

Storage, Sharing, Publication

Chapter 8

Storing personal data

In research, storage of personal data is one of the most common processing activities. Assuming you have a legal basis to store personal data, you then need to:

- Choose a storage medium that is GDPR-compliant and that provides a sufficient level of data protection;
- Take into account procedural and legal aspects, e.g., how will you handle the data once they are stored, and for how long will you store the data?

These aspects of storing personal data are discussed in this chapter.

8.0.1 Chapter summary

Where should I store personal data?

Use a medium that has been approved by your institution. If you work at Utrecht University, and your preferred storage medium is not included in the Storage Finder, then please contact RDM Support or your local data manager to find an alternative solution.

How to store personal data?

- Apply organisational and technical safeguards, e.g., restrict access, encrypt data, pseudonymise data, specify responsibilities, etc.
- Store (personal) data preferably in a structured, commonly used, machine-readable and interoperable format: others should be able to open, understand and work with your data.

For how long should I store personal data?

- Delete or fully anonymise personal data when they are no longer necessary, and preferably determine when you will do this in advance.
- In research, you can archive personal data that are necessary for validation purposes for a longer period of time, e.g., 10 years or longer.

8.1 Where should I store personal data?

On this page: storage, location, medium, yoda, o-drive, u-drive, usb stick, google drive, onedrive, teams, surfdrive, paper, security
Date of last review: 2022-06-02

If you work at Utrecht University (UU), you can find a suitable storage medium for digital research data via the Storage Finder. For personal data, select Sensitive or Critical (depending on the sensitivity of your data) under question 4 about Confidentiality.

Most storage media in this overview are suitable for storing personal data, either because they are controlled by UU (e.g., U- and O-drive, Beta File System) or because UU has a Data Processing Agreement in place with the storage supplier (e.g., Microsoft Office 365, Yoda).

Is your preferred storage medium not included in the storage finder? Contact RDM Support or your local data manager to find an alternative solution.

- Consider encrypting your data, especially when using **portable devices** (e.g, memory sticks, phones, dictaphones). Portable devices are also not suitable as back-up, due to bit rot and being easily lost.
- Physical personal data (e.g., paper questionnaires, informed consent forms) should be stored securely too, e.g. in a locked room, cabinet or drawer. You should also avoid leaving unsecured copies lying around (e.g., on a desk or printer).

Do not store research data containing personal data on public cloud services, e.g., Google Drive, Dropbox, OneDrive, Box, Mega, iDrive, iCloud, NextCloud, etc. These services are not (always) GDPR-compliant and/or may not offer sufficient data security. Moreover, UU does not have any formal agreements with these services, enabling them to use the data stored on their platforms for their own purposes.

8.2 How should I store personal data?

On this page: access control, accountability, interoperability, interoperable, separate, anonymise, pseudonymise, de-identify
Date of last review: 2022-06-02

Once you have chosen a suitable storage medium, you should act in accordance with the nature of your data as well, for example through:

- Controlling access: make sure that only the necessary people have the right kind of access (e.g. read/write) to the personal data, and remove their access when they do not longer need it (e.g. when someone leaves the research project).

- Specifying responsibilities, e.g. who is responsible for guarding access to the data on both the short and the long term? Make people aware of the confidential nature of the data. Tell them what to do in case of a data breach.
- Procedural arrangements, e.g. capture access conditions in agreements like the consortium agreement, data processing agreement or non-disclosure agreement.
- Storing different types of personal data in different places, e.g., research data should be stored separately from data subjects' contact details.
- Applying other safeguards where appropriate, e.g., encryption, pseudonymisation or anonymisation, etc.).

See Designing a GDPR-compliant research project for more tips.

Personal data should be stored in a **"structured, commonly used, machine-readable and interoperable format"** (rec. 68). In practice, this means that you should consider whether your files are structured and named in a logical way, use sustainable file formats, and provide understandable metadata so that others can interpret the data. You can read more about this in the RDM guide "Storing and preserving data".

8.3 For how long should I store personal data?

On this page: retention, storage period, duration, remove, delete
Date of last review: 2022-06-02

As per the GDPR, anyone processing personal data can only store those for as long as is necessary for prespecified purposes (art 5(e)). Afterwards, the personal data have to either be fully anonymised or deleted. However, there is an exemption for research data, as described below.

In research, we often see a division in different types of retention periods:

- If the personal data underpin a scientific publication, it is usually necessary to archive some personal data for **integrity and validation purposes** (art 5(e)), because they are part of the research data. At UU, any research data necessary for validation should be archived for at least 10 years (UU research data policy). If this includes personal data, they too should be archived. Importantly, this still means that you need to **protect** the personal data, and **limit** the personal data stored to the amount necessary for validation (art. 89)! This also implies that you should keep the documentation about the legal basis used (e.g., consent forms) during that time, so that you can demonstrate GDPR compliance.

- Specific retention periods may apply additionally to specific types of data. For example, in the Netherlands there are specific retention periods for medical data that range between 10 and 30 years at minimum.
- Personal data that were used for purposes other than answering your research question (e.g. contact information) should have their own retention policy: they should be removed or anonymised after the retention period (e.g. the research project) has ended.

If identification of the data subject is no longer needed for your (research) purposes, you do not need to keep storing the personal data just to comply with the GDPR, even if it means your data subjects cannot exercise their rights (art. 11).

For all types of data in your project (incl. to be archived research data), we recommend to formulate which data you will retain and for how long (for example in your Data Management Plan), and communicate the (possibly different) retention period(s) to data subjects. If you want to change the storage term you initially set and communicated for your personal data, please contact your privacy officer.

8.3.1 Deleting personal data

If you do not need personal data anymore, you must delete it, except when the data should be archived for validation purposes. When deleting data, it is important to make sure that there are no visible or hidden copies being left behind and that files cannot be recovered. The Storage Finder indicates how you can fully delete data on storage media within UU that are suitable for personal data. For your own file system, you can use software like BleachBit, BCWipe, DeleteOnClick, and Eraser to delete data.

Chapter 9

Sharing data with collaborators

On this page: share, transfer, collaborate, consortium, outside EU, EEA, security, legal basis, transparency, transparent, third-party transfer
Date of last review: 2022-09-22

This chapter addresses guidelines to take into account when you want to share personal data with collaborators outside of your own institution **during** your research project. For guidelines to share personal data after a research project, please refer to the chapter on Data sharing for reuse.

To be able to share personal data with external collaborators, you should:

1. Make sure you have a legal basis and inform data subjects

Make sure data subjects are well-informed about your intentions to share the data with collaborators. Include information in your information to data subjects on the identity of your collaborators, which data are shared with them and why, how, and for how long. Avoid using statements that preclude sharing such as "Your data will not be shared with anyone else".

Make sure you have a legal basis to share the data, e.g., informed consent or public interest. If you use consent, make sure that data subjects are aware that they are also providing consent to share their data with your collaborators.

Inform data subjects timely - before you start processing their data - and proactively - directly if possible.

2. Protect the personal data appropriately

Assess the risks of sharing the data and the measures you will take to mitigate those in your Data Management Plan, privacy scan, or if applicable, Data Protection Impact Assessment. This is especially important if you will share your data with collaborators outside of the European Economic Area.

Share only the data that the collaborator needs (data minimisation), for example by deleting unnecessary data, pseudonymising the data, and sharing only with those who need access to the data.

Make sure data subjects can still exercise their data subjects' rights. For example, if a data subject withdraws their consent, not only you, but also your collaborators will have to stop processing the data subject's personal data. It is important to make clear how you and them will do so.

3. Come to agreements with collaborators

In order to protect the personal data effectively, it is important to determine which role every collaborator has: controller or processor? And if there are multiple controllers, are they separate or joint controllers? For example, in many collaborative research projects (e.g., in consortia), there are multiple controllers that collectively determine why (e.g., research question) and how (e.g., methods) to process personal data. These parties are then joint controllers, and agreements need to be made in a joint controllers agreement. In any collaboration in which data are shared, you need to (art. 26):

Come to a formal agreement on:

The role of each party in the research project

Respective responsibilities in terms of data protection, such as informing data subjects and handling requests relating to data subjects' rights

Who is the main point of contact for data subjects

Communicate (the essence of) the agreement to data subjects.

Your privacy officer can help you draw up a valid agreement.

4. Pay special attention when sharing personal data outside the EU

If you share personal data with international collaborators (for example, with countries that have no adequacy decision, you may need to take additional measures. Usually, these measures include drawing up an agreement to make sure the other party is GDPR-compliant and uses the necessary security measures (if you haven't already done so). The exact type of agreement will depend on your specific situation: your privacy officer can help you choose and set up the right one. The flowchart below indicates conditions under which you can share data internationally. Note that they assume that you have taken sufficient safeguards to protect the personal data. To determine the possibilities of sharing data internationally in your project, we strongly advise you to consult with your privacy officer. In some cases a Data Transfer Impact Assessment may be required, which can take some effort.

5. Use a secure way to share the data

Granting access: It is preferable to grant a user access to an existing and safe infrastructure (e.g., add someone to a Yoda group or OneDrive folder), rather than physically sending the data elsewhere. This allows you to keep the data in one place, define specific access rights (read/write), have users authenticate, and easily revoke access to the data after your collaboration has ended. It is also a good idea to take measures to prevent the data from being copied elsewhere.

Transferring data: When it is absolutely necessary to transfer the files to a different location, you must do so securely. Researchers at Utrecht University can use SURF Filesender with encryption.

Chapter 10

Sharing data for reuse

On this page: publication, publish, share, transfer, open science, open data, FAIR data, reuse, reproducibility
Date of last review: 2023-03-09

In the context of Open Science, it is becoming more important to share data with other researchers, so that they can reproduce results and reuse the data to answer new research questions. This can be challenging when you work with personal data. Here, we list a few options for sharing personal data for reuse responsibly, and making datasets that contain personal data Findable, Accessible, Interoperable and Reusable (FAIR).

How you can make your dataset FAIR, from a privacy perspective, depends on which scenario applies to you:

- Your data are fully anonymised: you can publish them in a data repository.
- Your data cannot be fully anonymised, but you have a legal basis to share them for reuse: you can share them, but possibly with some restrictions.
- Your data cannot be fully anonymised, and you do not have a legal basis to share them: you cannot share them, but we discuss some alternatives.

If you are in doubt whether you can share personal data for reuse, please ask your privacy officer for help.

10.1 Sharing anonymised data

On this page: anonymous, publication, share, transfer, open science, open data, FAIR data
Date of last review: 2023-03-09

If (part of) the data are truly fully anonymous, they are not classified as personal data anymore: from a privacy perspective, you can publish this anonymised (part of the) dataset in a data repository without restrictions.

The data should indeed be fully anonymous. Here you can find how you know that your data are anonymous. In this preprint, you can find some examples of poorly anonymised (hence still personal) datasets in the field of psychology. It can be very difficult to fully anonymise personal data, so when in doubt, we recommend to always treat data as personal.

Just because data are no longer subject to the GDPR, it does not mean that there may not be other concerns for sharing data publicly. For example:

- publishing the data may not be *ethically* responsible when the data can be used to discriminate against a group of people.
- data may be someone else's intellectual property.

Publish your data in a data repository and include sufficient documentation and an open license to your dataset, to make your data FAIR. This way, others can find, access, understand and reuse your data.

10.2 Sharing personal data with a legal basis

On this page: pseudonymous, personal, sensitive, share, transfer, open science, reuse, access control, legal basis, legal ground, data sharing, transparency, transparent, inform, further processing, secondary use, secondary processing, safeguards, protection, FAIR data
Date of last review: 2023-03-09

If you cannot fully anonymise your data, they are still considered personal data. In order to share personal data for reuse, you therefore need to consider the following steps:

1. Be transparent in your information to data subjects
2. Make sure you have a legal basis
3. Protect the data while sharing
4. Make your personal data FAIR

If you are in doubt whether you can share personal data for reuse, please ask your privacy officer for help. If you cannot share the personal data for reuse, there are still alternatives you can apply to make (characteristics of) your data useful to others.

10.2.1 1. Be transparent

Irrespective of the legal basis you use to share personal data, data subjects must be informed about any reuse of their data. This allows them to exercise their rights, such as the right to object (if you use public interest) or to withdraw their consent (if you use consent). **If data subjects haven't been informed that you will share their data, you cannot share their data**: you have not fulfilled your transparency obligation!

Before the start of your project

Include the intention of sharing data in your information to data subjects, how you plan to keep them informed, and how they can exercise their rights. Avoid language that precludes sharing, such as "your data will remain strictly confidential", and "your data will only be shared with members of the research team".

If it is not possible to identify the specific data subject that objected or withdrew consent within the dataset, without additional information provided by the data subject themselves, data subjects can simply not exercise those rights anymore. Let data subjects know about this!

At the time of data sharing

If you can still identify the data subjects in your dataset at the time of data sharing (e.g., if you still have a keyfile and/or contact information), inform the data subjects specifically about the data sharing process, using appropriate channels such as email (art. 12, art. 14): which data are shared, with whom exactly, for which purposes, under which restrictions, and how can data subjects object or withdraw consent?

If you cannot identify data subjects in the dataset at the time of data sharing (e.g., there is no keyfile/contact information anymore, but the data are not anonymous), inform them indirectly on how their data are being (re)used and if/how they can exercise their rights, via channels that are easily accessible, for example through a project website, newsletter, mailing list, etc.

In most cases, the original owner (controller) of the data is responsible for informing data subjects and handling requests related to data subjects' rights, unless otherwise agreed.

10.2.2 2. Make sure you have a legal basis

When you share personal data with another organisation for their own specified reuse, the recipient will likely become a new controller of the personal data. This means that both you and the recipient need a valid legal reason to share (you, the owner) and (re)use (recipient) the personal data.

For the **original owner**, there are multiple possibilities to rely on to share the data:

Further processing for research purposes

This is a derogation in the GDPR that enables personal data to be further processed (e.g., shared) for *research purposes*, without requiring a new legal basis, as long as sufficient safeguards are in place to protect the data (e.g., pseudonymisation, access control, data transfer agreement, etc, art. 5(1)(b), art. 89). If the data are not shared for research purposes, then this derogation does not apply. **Note**: There is ongoing discussion whether you can rely on this derogation if you used consent to collect the data: sharing that falls outside of the scope of the original consent, may not meet the specificity criterium and may not be fair to data subjects. You can rely on it, however, if you used another legal basis to collect the data (e.g., public interest, legitimate interest).

Consent for data sharing

This entails asking explicit consent to share data with others for reuse for specified purposes, before you collect the data. In this guide you can find more information about that.

An advantage of this approach is that it gives data subjects a lot of control, and reuse does not have to be limited to scientific research only (as it is with further processing).

A limitation of this approach is that consent has to be specific in order to be valid. Thus, consent for data sharing is only legitimate when you *additionally* inform data subjects about the *specific* sharing right before you share the data (e.g., with whom specifically will the data be shared and why?), so that data subjects can still withdraw their data sharing consent.

Public or legitimate interest

Public or legitimate interest could in principle also be used as a legal bases to share personal data, when sharing the data is necessary and proportional and it does not override the interests of the data subjects (a privacy scan is a good way to assess that).

For the recipient, in most cases the legal basis for reusing the received data is public interest (when reused for research purposes), although legitimate interest (when reused for non-research purposes) and consent (if the recipient can themselves obtain consent from the data subjects) are also possible. Using public (and legitimate) interest requires the recipient to assess the risks for data subjects against the benefits of using the data for their purposes (a privacy scan is a good way to do that). This is necessary because the recipient will become a new controller and therefore also has to treat the personal data in a fair, transparent and lawful way. The recipient is usually also bound by the restrictions set forth by the original owner, which usually happens through a data transfer agreement or custom license (e.g., use safeguards to protect the data, do not share the data any further, only use the data for the specified purposes, etc.).

If you want to share or reuse special categories of personal data, you may still need explicit consent, except when the data subject had made their data publicly available themselves, or when obtaining consent would involve an unreasonable amount of effort.

10.2.3 3. Protect the data while sharing

Unless you have a legal basis to make personal data publicly available, you should aim to protect the personal data also while sharing them. For example:

- Do not share more data than needed; pseudonymise the data as much as possible.
- Put in place an agreement that forces recipients to treat the data confidentially and that clarifies each party's responsibilities. This is especially important if the data will be shared outside the European Economic Area with countries that do not have an adequate level of data protection.
- Share the data safely, for example by giving access via a secure storage environment, or encrypting the data before transferring them.
- Always follow the restrictions that you communicated to data subjects.

10.2.4 4. Make your data FAIR

Personal data or not, you can always make your data Findable, Accessible, Interoperable and Reusable:

Findable

Publish your metadata and documentation in a data repository that assigns a persistent identifier to the dataset. Depending on your situation, you may be able to deposit the data there as well (under restricted access).

Accessible

Clearly specify if and how others can access your dataset and make that information publicly available. Some studies have set up a data access protocol in which this is made clear (e.g., data are accessible after signing an agreement, writing a research proposal, helping to collect new data, etc.). You can find an example here.

Interoperable

Structure and document your data so that they are easily understandable for humans and machines (see our FAIR guide).

Reusable

If you only deposited metadata and documentation, add an open license to the dataset (e.g., CC0 or CC BY 4.0). If you deposited the personal data in the data repository as well, there will usually be custom terms of use such as the data access protocol mentioned under Accessible.

10.3 Alternatives to sharing personal data

On this page: metadata, documentation, information, publication, share, transfer, open science, FAIR data, reproducibility
Date of last review: 2023-03-09

10.3.1 Publish metadata and documentation

Even if you cannot share/publish the data, you can still publish non-sensitive metadata and documentation surrounding your research project. This allows your dataset and documentation to be findable, citable, and in some cases even reusable (one person's metadata is another person's data!). In order to make the dataset FAIR, you should include a note on the access restrictions of the dataset and choose a good data repository. Knowing that your dataset exists can sometimes already be useful information, even when the data are not accessible for others. For an example, please refer to the use case about the Open Science Monitor.

10.3.2 Use other techniques and strategies to enable reuse

There are also more technical alternatives to transferring personal data to others:

- Use solutions that allow others to run analyses on your data, without ever needing access to those data (remote data science, see the Secure computing chapter).
- Create a synthetic dataset that others can use to reproduce trends or explore the data.
- Only allow differentially private algorithms to query your dataset.
- Publish aggregated (anonymous) data which may still be useful for others (e.g., group-level statistics).

Techniques & Tools

Chapter 11

Pseudonymisation & Anonymisation

On this page: anonymous, pseudonymous, deidentification, safeguard, protection measure, sdc, statistical disclosure control
Date of last review: 2023-05-02

Pseudonymisation and anonymisation are both ways to make personal data less easily linkable to individual data subjects: they are methods to **de-identify** personal data. Importantly, whereas anonymisation results in non-personal data that are not subject to the GDPR anymore, pseudonymised data are **still personal data**. It is therefore important to understand the difference between the two, and to estimate when your data are indeed fully anonymous.

Any operation that you do up until the personal data are anonymised - including the anonymisation itself - is still subject to the GDPR. So even if you can anonymise your data later, you still need to comply with the GDPR for everything you do beforehand (e.g., collecting, analysing, sharing, etc.).

In this chapter, we:

1. Explain what pseudonymisation and anonymisation mean.
2. Present a step-by-step workflow to de-identify personal data.
3. List a number of techniques that you can use to de-identify personal data.

Finally, we list some resources for further reading.

11.1 What are pseudonymisation and anonymisation?

On this page: anonymous, pseudonymous, deidentification, safeguard, protection measure, identifiable, sdc, statistical disclosure control, disclosure risk
Date of last review: 2023-05-02

11.1.1 Pseudonymisation

Pseudonymisation is a safeguard that reduces the linkability of your data to your data subjects (rec. 28). It means that you de-identify the data in such a way that they can no longer lead to identification *without additional information* (art. 4(5)). In theory, removing this additional information should lead to anonymised data.

Pseudonymisation is often interpreted as replacing direct identifiers (e.g., names) with pseudonyms, and storing the link between the identifiers and the pseudonyms in a key file, separated from the research data. While this is a good practice (it makes sure that data are not directly identifiable anymore), this interpretation of pseudonymisation does not take into account indirectly identifiable information, and thus does not necessarily fulfil the GDPR's definition of pseudonymisation!

Pseudonymous data are still **personal data** and thus subject to the GDPR. This is because the de-identification is *reversible*: identifying data subjects is still possible, just more difficult. This means that in order to use pseudonymous data, you still need to comply to all the rules in the GDPR.

11.1.2 Anonymisation

Anonymisation is a de-identification process that results in data that are "rendered anonymous in such a manner that the data subject is not or no longer identifiable" (rec. 26), neither directly nor indirectly, and by no one, including you. When data are anonymised, they are no longer personal data, and thus no longer subject to the GDPR. Note, however, that **everything you do *before* the data are anonymised** (including the anonymisation itself) *is* subject to the GDPR!

Anonymisation is very difficult to accomplish in practice! This video nicely illustrates why.

11.1.3 The identifiability spectrum

The relationship between (identifiable) personal data, pseudonymous data and anonymous data should be seen as lying on a spectrum. The more de-identified the data are, the closer they are to anonymous data and the lower the risk of re-identification. The visual guide below nicely illustrates this:

A VISUAL GUIDE TO PRACTICAL DATA DE-IDENTIFICATION

Produced by
FUTURE OF PRIVACY FORUM
FPF.org

What do scientists, regulators and lawyers mean when they talk about de-identification? How does anonymous data differ from pseudonymous or de-identified information? Data identifiability is not binary. Data lies on a spectrum with multiple shades of identifiability.

This is a primer on how to distinguish different categories of data.

DEGREES OF IDENTIFIABILITY

Information containing direct and indirect identifiers.

PSEUDONYMOUS DATA

Information from which direct identifiers have been eliminated or transformed, but indirect identifiers remain intact.

DE-IDENTIFIED DATA

Direct and known indirect identifiers have been removed or manipulated to break the linkage to real world identities.

ANONYMOUS DATA

Direct and indirect identifiers have been removed or manipulated together with mathematical and technical guarantees to prevent re-identification.

	DEGREES OF IDENTIFIABILITY			PSEUDONYMOUS DATA			DE-IDENTIFIED DATA		ANONYMOUS DATA	
	EXPLICITLY PERSONAL	POTENTIALLY IDENTIFIABLE	NOT READILY IDENTIFIABLE	KEY CODED	PSEUDONYMOUS	PROTECTED PSEUDONYMOUS	DE-IDENTIFIED	PROTECTED DE-IDENTIFIED	ANONYMOUS	AGGREGATED ANONYMOUS
DIRECT IDENTIFIERS Data that identifies a person without additional information or by linking to information in the public domain (e.g., name, SSN)	INTACT	PARTIALLY MASKED	PARTIALLY MASKED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED
INDIRECT IDENTIFIERS Data that identifies an individual indirectly. Helps connect pieces of information until an individual can be singled out (e.g., DOB, gender)	INTACT	INTACT	INTACT	INTACT	INTACT	INTACT	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED	ELIMINATED or TRANSFORMED
SAFEGUARDS and CONTROLS Technical, organizational and legal controls preventing employees, researchers or other third parties from re-identifying individuals	NOT RELEVANT due to nature of data	LIMITED or NONE IN PLACE	CONTROLS IN PLACE	CONTROLS IN PLACE	LIMITED or NONE IN PLACE	CONTROLS IN PLACE	LIMITED or NONE IN PLACE	CONTROLS IN PLACE	NOT RELEVANT due to nature of data	NOT RELEVANT due to high degree of data aggregation
SELECTED EXAMPLES	Name, address, phone number, SSN, government-issued ID (e.g., Jane Smith, 123 Main Street, 555-555-5555)	Unique device ID, license plate, medical record number, cookie, IP address (e.g., MAC address 68:A8:6D:35:65:03)	Same as Potentially Identifiable except data are also protected by safeguards and controls (e.g., hashed MAC addresses & legal representations)	Clinical or research datasets where only curator retains key (e.g., Jane Smith, diabetes, HgB 15.1 g/dl = Csrk123)	Unique, artificial pseudonyms replace direct identifiers (e.g., HIPAA Limited Datasets, John Doe = 5L7T LX619Z) (unique sequence not used anywhere else)	Same as Pseudonymous, except data are also protected by safeguards and controls	Data are suppressed, generalized, perturbed, swapped, etc. (e.g., GPA: 3.2 = 3.0-3.5, gender: female = gender: male)	Same as De-Identified, except data are also protected by safeguards and controls	For example, noise is calibrated to a data set to hide whether an individual is present or not (differential privacy)	Very highly aggregated data (e.g., statistical data, census data, or population data that 52.6% of Washington, DC residents are women)

If the image does not show correctly, view it online

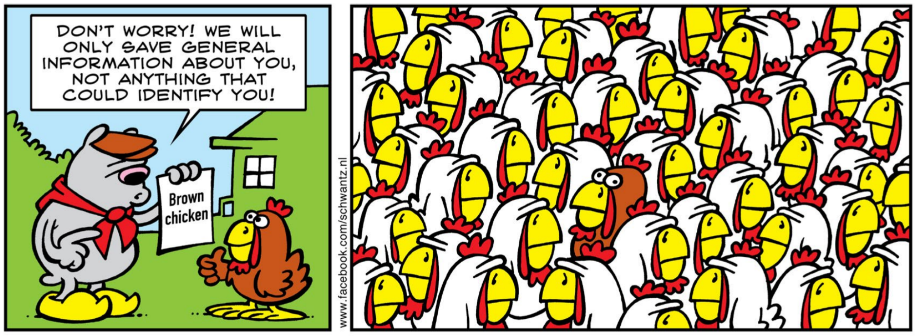
11.1.4 When are data anonymous?

Your data can be considered anonymous if data subjects can only be re-identified with an *unreasonable* amount of effort, i.e., taking into account the costs, required time and technology, and future technological developments (rec. 26).

Basically, your data are **not** anonymous (personal) when they comply with *any* of the characteristics of personal data:

- There is directly identifiable information (e.g., name, email address, social security number, etc.).
- Data subjects can be singled out (i.e., you can tell one data subject from another within a known group of data subjects).
- It is possible to identify data subjects by linking records ("mosaic effect"), either within your own database or when using other data sources.
- It is possible to identify a data subject by inferring information about them (e.g., infer a disease by the variable "medication"), either within your own database or when using other data sources.
- It is possible to reverse the de-identification.

Whether data can be seen as anonymous strongly depends on the context of your research and how much information is available about the data subjects.


DON'T WORRY! WE WILL ONLY SAVE GENERAL INFORMATION ABOUT YOU, NOT ANYTHING THAT COULD IDENTIFY YOU!
Brown chicken
www.facebook.com/schwantz.nl

When collaborating with research data centres, such as the Statistics Netherlands (Centraal Bureau voor de Statistiek, CBS), often output checking guidelines are used to determine the risk of identification resulting from the analysis output of sensitive data.

11.1.5 Alternatives to anonymisation

Anonymisation is not the only solution. The best way to protect data subjects' privacy is to only collect/process their personal data if necessary (minimisation).

Additionally, in many cases, full anonymisation is not even possible or desirable, for example if it results in too much information loss or incorrect inferences.

If you cannot anonymise the data, there are always other ways in which you can protect the data, such as:

- De-identifying (pseudonymising) the data to the extent you can.
- Controlling access to the data, for example using user agreements, authentication, encryption, secure analysis environments, etc.
- Creating a synthetic version of your dataset to share with others.

11.2 Step-by-step de-identification

On this page: anonymous, pseudonymous, step-by-step, workflow, deidentification, safeguard, protection measure
Date of last review: 2023-05-02

Below is a step-by-step workflow that you can use to de-identify your data. Alternatively, you could also use this de-identification plan template to plan and document your de-identification steps. Whether or not the de-identification results in a pseudonymised or an anonymised dataset is highly dependent on the characteristics of the dataset and the context in which it was obtained.

1. Perform the de-identification in a **safe storage or processing environment**: remember that you are working with personal data, and as long as the data are not anonymous, they will be subject to the GDPR!
2. Identify any **potentially identifying information** in your data.
3. Assess whether you need to **collect this information at all**. For example:

    a. Do you really need IP addresses in your survey data?
    b. Do you really need to record audio or video?
    c. Do you really need a consent form with a name, contact information, and signature on it?
    d. Replace names with pseudonyms in filenames and within the data where possible.

4. If you do not need **directly identifying information** to answer your research question, but you do need it to, for example, contact data subjects:

    a. Separate directly identifying information from the research data.
    b. Use pseudonyms or hashes to refer to individuals instead of names.
    c. Create a keyfile to link the pseudonyms to the names.
    d. Store the directly identifiable information and the keyfile in a separate location from the research data and/or in encrypted form.

5. Consider which types of information may lead to **indirect identification**, such as demographic information (age, education, occupation, etc.), geolocation, specific dates, medical conditions, unique personal characteristics, open text responses, etc.
6. **De-identify** the directly and indirectly identifiable data using (a selection of) the techniques described on the next page.

   a. Before you start, save a copy of the raw, untouched dataset, in case anything in the process goes wrong.
   b. Document the steps you took, for example in a programming script or README file, which always accompanies the data.
   c. Whether you can delete the raw (non-pseudonymised) version of the dataset, depends on whether it needs to be preserved for verification purposes. Specific restrictions may also apply if the Dutch Medical Research Involving Human Subjects Act (WMO) and/or Good Clinical Practice apply to your research.

7. **Treat the data according to their sensitivity**. If the data are not fully anonymised, they are pseudonymous and thus still need to be handled according to the GDPR guidelines!

How de-identified is de-identified enough? You can read more about this in the chapter Statistical approaches to privacy.

11.3 De-identification techniques

On this page: anonymous, pseudonymous, deidentification, safeguard, protection measure, technique, anonymisation method, privacy-preserving, privacy-enhancing, sdc, statistical disclosure control, disclosure risk
Date of last review: 2023-05-02

Below is a list of techniques you can apply to your data to de-identify your dataset so that it results in a pseudonymised, or possibly even anonymised dataset. Bear in mind that applying these will always result in loss of information, so ask yourself how useful your dataset will still be after de-identification.

The techniques are:

- Suppression
- Generalisation
- Replacement
- Top- and bottom coding
- Adding noise
- Permutation

Statistical Disclosure Control (SDC) The below de-identification methods are sometimes also referred to as methods to apply Statistical Disclosure Control (SDC). You will most likely encounter SDC when you collaborate with a research data centre such as Statistics Netherlands (Centraal Bureau voor de Statistiek, CBS).

11.3.0.1 Suppression

Suppression (sometimes called "masking") basically means removing variables, (parts of) values, or entire entries that you do not need from your dataset. Examples of data that you could consider removing:

- Name and contact information
- (Parts of) address
- Date, such as birthdate or participation date
- Social security number/Burgerservicenummer (BSN). NB. In the Netherlands, you are not allowed to use BSN in research at all!
- Medical record number
- IP address
- Facial features from neuroimaging data
- Automatically generated metadata such as GPS data in an image, author in a document, etc.
- Participants that form extreme outliers or are too unique

11.3.0.2 Generalisation

Generalisation (also sometimes called abstraction, binning, aggregation, or categorisation) reduces the granularity of the data so that data subjects are less easily singled out. It can be applied to both qualitative (e.g., interview notes) and quantitative data (e.g., variables in a dataset). Here are some examples:

- Recoding date of birth into age.
- Categorising age into age groups.
- Recoding rare categories as "other" or as missing values.
- Replacing address with the name of a neighbourhood or town.
- Generalising specific persons in text into broader categories, e.g., "mother" to "[woman]", "Bob" to "[colleague]".
- Generalising specific locations into more general places, e.g., "Utrecht" to "[home town]", or from point coordinates to larger geographical areas (e.g., polygon or linear features).
- Coding open-ended responses into categories of responses, or as "responded" vs. "not responded".

11.3.0.3 Replacement

In this case, you replace sensitive details with non-sensitive ones, which are usually less informative, for example:

- Replacing directly identifying information that you do need with pseudonyms. When doing this, always store the key file securely and separately from the research data (e.g., use access control, encryption). If you do not need the links with direct identifiers anymore, remove the keyfile or replace the pseudonyms with random identifiers without saving the key.
  A good pseudonym:
  Is not meaningful with respect to the data subjects: a random (unique) number or string is better than a code that contains parts of personal information, because the latter may reveal details about data subjects.
  Is managed securely, for example by appointing someone to be responsible for managing access to the keyfile.
  Can be a simple number, random number, cryptographic hash function, text string, etc. (read more).
- Replacing identifiable text with "[redacted]". When redacting changes in-text, never just blank out the identifying value, always put a placeholder or pseudonym there, e.g., in `[square brackets]` or `<seg>`segments`</seg>`.
- Replacing unique values with a summary statistic, e.g., the mean.
- Rounding values, making the data less precise.
- Replacing one or multiple variables with a hash.
  What is hashing?
  Hashing is a way of obscuring data with a string of seemingly random characters with a fixed length. It can be used to create a"hashed" pseudonym, or to replace multiple variables with one unique value. There are many hash functions which all have their own strength. It is usually quite difficult to reverse the hashing process, except if an attacker has knowledge about the type of information that was masked through hashing (e.g., for the MD5 algorithm, there are many lookup tables that can reverse common hashes). To prevent reversal, cryptographic hashing techniques add a "salt", i.e., a random number or string, to the hash (the result is called a"digest"). If the "salt" is kept confidential or is removed (similar to a keyfile), it is almost impossible to reverse the hashing process.

11.3.0.4 Top- and bottom-coding

Top- and bottom-coding are mostly useful for quantitative datasets that have some unique extreme values. It means that you set a maximum or minimum and recode all higher or lower values to that minimum or maximum. For example, you can top-code a variable "income" so that all incomes over €80.000 are set to €80.000. This does distort the distribution, but leaves a large part of the data intact.

11.3.0.5 Adding noise

Adding noise to data obfuscates sensitive details. It is mostly applied to quantitative datasets, but can also apply to other types of data. For example:

- Adding half a standard deviation to a variable.
- Multiplying a variable by a random number.
- Applying Differential Privacy guarantees to an algorithm.
- Blurring (pixelating) images and videos.
- Voice alteration in audio.

11.3.0.6 Permutation

Permutation means swapping values between data subjects, so that it becomes more difficult to link information belonging to one data subject together. This will keep the distribution and summary statistics constant, but change correlations between variables, making some statistical analyses more difficult or impossible.

11.4 Tools and further reading

On this page: anonymous, pseudonymous, deidentification, safeguard, protection measure, tool, resource, reading material
Date of last review: 2023-05-02

You can find a selection of **de-identification tools** in this GitHub repository.

For further reading, we compiled a reading list on this topic in our publicly accessible Zotero library.

We can recommend:

- 10 misunderstandings related to anonymisation.
- Risk management for research data about people.
- CESSDA's Data Management Expert Guide on Anonymisation.
- Anonymisation: managing data protection risk, code of practice.
- Privacy protection in the era of open science.
- An in-depth overview of anonymisation techniques and tools
- Statistics Netherlands (CBS) on techniques used to de-identify sensitive data.

Chapter 12

Statistical approaches to de-identification

In order to protect datasets that contain personal or otherwise sensitive data, there is an increasing number of statistical approaches to de-identification, which to some extent quantify how identifiable data are after de-identification.

In this chapter, we discuss the following approaches, as these are the most widely used approaches:

- K-anonymity
- L-diversity
- T-closeness
- Differential privacy

These approaches (or: privacy models) are not yet much used in research practice, because they come with some disadvantages and require resources and/or expertise to be applied and interpreted correctly. However, they are used in many de-identification tools and are useful to detect specific sensitivities in (tabular) datasets. For those reasons, the techniques are introduced in this chapter.

12.1 K-anonymity, l-diversity and t-closeness

On this page: k-anonymous, l-diverse, t-close, privacy model, quantifying privacy, key attribute, sensitive attribute, quasi-identifier
Date of last review: 2023-05-30

K-anonymity, L-diversity and T-closeness are statistical approaches that quantify the level of identifiability within a tabular dataset, especially when variables within that dataset are combined. They are complementary approaches: a dataset can be k-anonymous, L-diverse and T-close, where k, L and T all represent a number.

12.1.1 Identifiers, quasi-identifiers, and sensitive attributes

Privacy models like k-anonymity, L-diversity and T-closeness distinguish between 3 types of variables in a dataset:

- **Identifiers** (also known as key attributes): direct identifiers such as names, student numbers, email addresses, etc. These variables should in principle not be collected at all, or removed from the dataset if they are not necessary for your research project.
- **Quasi-identifiers**: indirect identifiers that can lead to identification when combined with other quasi-identifiers in the dataset or external information. These are often demographic variables like age, sex, place of residence, etc., but could also be something entirely different like physical characteristics, timestamps, etc. In general, quasi-identifiers are usually variables that are likely to be known to someone in the outside world.

- **Sensitive attributes**: variables of interest which should be protected, and which cannot be changed, because they are the main outcome variables. For example, it can be Medical condition in a healthcare dataset, or Income in a financial dataset.

It is important to correctly categorise the variables in your dataset as any of these variable types if you want to apply k-anonymity, l-diversity and t-closeness, because they will determine how the dataset will be de-identified.

12.1.2 How it works

12.1.2.1 K-anonymity

K-anonymity ensures that each individual in a dataset cannot be distinguished from at least k-1 other individuals with respect to the quasi-identifiers in the dataset. This is done through generalisation, suppression and sometimes top- and bottom-coding. Applying k-anonymity makes it more difficult for an attacker to re-identify specific individuals in the dataset. It protects against singling out and, to some extent, the Mosaic effect.

Table 12.1: Original dataset

Nr	Age	Sex	City	Disease
1	**16**	Male	Rotterdam	Viral infection
2	**18**	Male	Rotterdam	Heart-related
3	**19**	Male	Rotterdam	Cancer
4	**22**	Female	Rotterdam	Viral infection
5	**22**	Male	Zwolle	No illness
6	**23**	Male	Zwolle	Tuberculosis
7	**24**	Male	Zwolle	Heart-related
8	**25**	Female	Utrecht	Cancer
9	**26**	Female	Rotterdam	Heart-related
10	**28**	Female	Utrecht	Tuberculosis

Table 12.2: 2-anonymous dataset

Nr	Age	Sex	City	Disease
1	**=< 20**	Male	Rotterdam	Viral infection
2	**=< 20**	Male	Rotterdam	Heart-related
3	**=< 20**	Male	Rotterdam	Cancer
4	**20-30**	Female	Rotterdam	Viral infection
5	**20-30**	Male	Zwolle	No illness
6	**20-30**	Male	Zwolle	Tuberculosis
7	**20-30**	Male	Zwolle	Heart-related
8	**20-30**	Female	Utrecht	Cancer
9	**20-30**	Female	Rotterdam	Heart-related
10	**20-30**	Female	Utrecht	Tuberculosis

Colours indicate an 'equivalence class' of quasi-identifers

To make a dataset k-anonymous, you must first identify which variables in the dataset are identifiers, quasi-identifiers, and sensitive attributes. In the example above, Age, Sex and City are quasi-identifiers and Disease is the sensitive attribute. Next, you should set a value for k. If we choose a k of 2, every row in the example dataset should have the same combination of Age, Sex and City as at least 1 other row in the dataset. Finally, you aggregate the dataset so that every combination of quasi-identifiers occurs at least k times. In the example, this was done by generalising Age into age categories, but there may also be other ways to reach 2-anonymity in this dataset.

There is no single value for k which you should always choose. The higher the k, the more difficult it will be to identify someone, but likely your dataset will also become less granular and perhaps less informative. The value of k will be highly dependent on what you communicated to data subjects (e.g., you may have promised a certain k) and the risk of identification that you are willing to accept.

The below video gives an example on how k-anonymity can work in practice:

12.1.2.2 L-diversity

L-diversity is an extension to k-anonymity that ensures that there is sufficient variation in a *sensitive attribute*. This is important, because if all individuals in a (subset of a) dataset have the same value for the sensitive attribute, there is still a risk of *inference*. For example, in the below 2-anonymous dataset, you can infer that any female from Rotterdam between 20 and 30 who participated had a viral infection ("homogeneity attack"). Similarly, if you know that your 25-year old female neighbour from Utrecht participated in this study, you learn that she suffers from cancer ("background knowledge attack").

K-anonymity does not protect against such homogeneity and background knowledge attacks. Therefore, L-diversity proposes that there should be at least L different values for the sensitive attribute per combination of quasi-identifiers. In the example above, if we choose an L of 2, that means that for each combination of Age, Sex and City, there are at least 2 distinct diseases. In the example, we suppressed City for these homogeneous cases, so that all females between 20 and 30 years old can either have cancer or a viral infection.

Like k-anonymity, there is no perfect value of L, although it is usually less or equal to k and more than 1.

The below video explains the concept of L-diversity using an example:

12.1.2.3 T-closeness

T-closeness ensures that the distribution of a *sensitive attribute* within a generalisation of a *quasi-identifier* is close to the distribution of the sensitive attribute

Table 12.3: 2-anonymous dataset

Nr	Age	Sex	City	Disease
1	=< **20**	Male	Rotterdam	Viral infection
2	=< **20**	Male	Rotterdam	Heart-related
3	=< **20**	Male	Rotterdam	Cancer
4	**20-30**	Female	Rotterdam	Viral infection
5	**20-30**	Male	Zwolle	No illness
6	**20-30**	Male	Zwolle	Tuberculosis
7	**20-30**	Male	Zwolle	Heart-related
8	**20-30**	Female	Utrecht	Cancer
9	**20-30**	Female	Rotterdam	Viral infection
10	**20-30**	Female	Utrecht	Cancer

Colours indicate an 'equivalence class' of quasi-identifers

Table 12.4: 2-anonymous 2-diverse dataset

Nr	Age	Sex	City	Disease
1	=< **20**	Male	Rotterdam	Viral infection
2	=< **20**	Male	Rotterdam	Heart-related
3	=< **20**	Male	Rotterdam	Cancer
4	**20-30**	Female	*	Viral infection
5	**20-30**	Male	Zwolle	No illness
6	**20-30**	Male	Zwolle	Tuberculosis
7	**20-30**	Male	Zwolle	Heart-related
8	**20-30**	Female	*	Cancer
9	**20-30**	Female	*	Viral infection
10	**20-30**	Female	*	Cancer

Colours indicate an 'equivalence class' of quasi-identifers and sensitive attributes in the entire dataset. In other words, it ensures that the sensitive attribute is not skewed towards a specific value within a group of similar individuals, which could potentially be used to re-identify someone. For example, if a dataset contains information on Age (quasi-identifier), Sex (quasi-identifier), and Income (sensitive attribute), and t-closeness is applied with a value of t = 0.1, then for each combination of Age and Sex, the distribution of income must be within 10% of the distribution of income in the entire dataset.

T-closeness can get complicated quite fast. If you're curious to know how it works, the below video explains the concept of t-closeness using an example:

12.1.3 When to use

K-anonymity, L-diversity and t-closeness are usually applied to de-identify tabular datasets, before being shared. They are also most suitable for relatively large datasets (i.e., containing a large number of individuals), as more details (utility) are likely to be retained in such datasets (source).

12.1.4 Implications for research

- It is very easy to lose a lot of the (granularity of the) data when satisfying the k-, L- or T-criteria: the higher the criteria, the lower the risk of re-identification, but the more information you lose. The balance between privacy and utility is therefore very important to take into consideration when applying these privacy models.
- The more variables (quasi-identifiers), the larger the dataset and the more outliers there are in the dataset, the more difficult de-identification will be without losing too much information (as shown here).

- If a dataset is k-anonymous, L-diverse or T-close, that does not mean that the dataset is also considered anonymous under the GDPR. The degree of anonymity after applying these approaches depends entirely on your own choices in terms of k, L or T, in terms of the variables that you included, and on the context of your dataset. For example, if you failed to include a quasi-identifier in k-anonymising your dataset, your dataset is in reality not k-anonymous.

12.1.5 Further reading

- Read the original papers about k-anonymity (Sweeney, 2002, pdf), L-diversity (Machanavajjhala et al., 2007) and t-closeness (Li et al., 2007, pdf).

- In response to k-anonymity, L-diversity and T-closeness, other approaches have been formulated as well, such as k-Map, delta-presence ( ), beta-likeness ( ) and delta-disclosure ( ). We do not go into those here, but Damien Desfontaines's blogposts and this thesis (pdf) discuss some of these approaches in more detail.
- On this page and here you can find more information on the different attacks that can potentially be committed to datasets.

12.2 Differential privacy

On this page: differential privacy, dp, epsilon, quantifying privacy, privacy-utility spectrum Date of last review: 2023-05-30

Differential privacy (DP) is a mathematical technique to prevent the disclosure of personal information by adding statistical noise to queries. In its original form, the data are stored on a secure server. Researchers can then, without access to the data, pose a query (i.e., an analysis request), and obtain the result with a certain amount of statistical noise added to the output. The amount of noise is tweaked such that adding or removing an individual to or from the dataset alters the result of the query by a limited, predetermined amount.

The below video explains the basic premise of differential privacy using a very basic example:

Differential privacy is generally not added to a dataset itself, but to the **result of a query** (e.g., average income in a dataset), although some forms exist in which the dataset itself is altered; this is called *local* differential privacy.

12.2.1 How it works

Imagine a dataset with peoples' incomes, and we want to know the average income in the data (i.e. our query). When applying differential privacy, we want to ensure that we cannot determine whether an individual is in the dataset or not. In theory, we can do this by adding or removing any single individual to or from our dataset and recalculating the average income in the modified dataset. Importantly, additions to the dataset should all be individuals that *could be* in the dataset, and not only those that already are. Depending on the privacy requirements (i.e., what amount of privacy leakage is acceptable) and the sensitivity of the query (the effect of changing one individual on the average income), more or less statistical noise is added to the result.

The most common definition of differential privacy is "epsilon-differential privacy", epsilon being the "**privacy budget**". In this definition, we consider two scenarios: the original dataset and a modified dataset, where we substitute the person with the highest income with the person with the lowest income. A certain amount of noise is added, so that the probabilities of obtaining a certain result in the original and the modified datasets do not differ more than Epsilon. Epsilon (i.e., the privacy budget) is a predetermined value: the lower epsilon, the higher the level of privacy, and therefore the more noise is added to the query.

Differential privacy can also protect *groups* of people by adding or subtracting groups instead of individuals. This can for example be useful when you want to protect a household as a whole, and not only the individuals in it.

Setting a lower epsilon results in more noise in the results that you release. This simply implies that it becomes more difficult to distinguish whether individuals were or were not in your dataset. However, it does **not** necessarily imply that a certain level of epsilon will result in fully anonymous data. Whether a dataset can be viewed as anonymous or not under the GDPR will always be context- and dataset-dependent.

12.2.2 Implications for research

- The advantage of differential privacy is that privacy risk can be mathematically quantified (through epsilon). With many other techniques, it can be hard to determine exactly what the privacy guarantees are, especially with the presence of external information.
- As with the other statistical approaches, there is a trade-off between more privacy or more utility: the lower the privacy budget, the more privacy is retained, but also the less accurate (more noisy) individual values will be.
- The privacy budget should be determined in advance and can only be spent once: every query will lower the available privacy budget by epsilon, and when the privacy budget reaches zero, no more queries can be done. Thus, **you cannot indefinitely reuse the data and still preserve privacy**.
- A disadvantage of differential privacy is that the concept is not completely trivial for the non-expert and in some cases, this has resulted in violated privacy guarantees. There is **no absolute right way to set the privacy budget** and no framework to decide which value of epsilon should be set for what kind of data. Thus, it can be hard for researchers to justify using a particular value of epsilon.

12.2.3 When to use

The implementation of differential privacy is a very technical (and tricky) endeavour. Setting up a differentially private server that only outputs differentially private results for any query is currently practically impossible. Thus, if you are not an expert on statistics and/or differential privacy, we recommend reaching out to someone who is, and to use Differential Privacy in addition to other privacy-enhancing techniques, such as pseudonymisation.

For now, differential Privacy is mostly used in combination with synthetic data, or with simple repetitive queries. Widespread use of differential privacy might become safer and/or easier over time as implementations are tested more thoroughly on real-world datasets.

12.2.4 Further reading

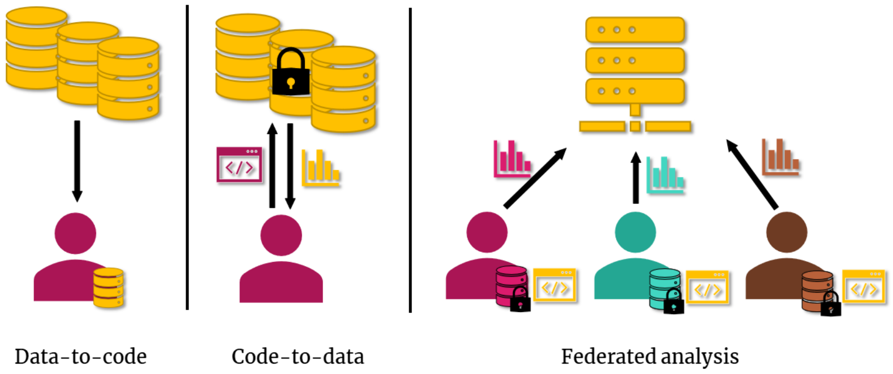
- Damien Desfontaines has written a number of easy to read blogposts about differential privacy, including this one explaining the basic premise of differential privacy.
- The online book Programming Differential Privacy explains in detail how differential privacy works.
- Or read Differential privacy: a primer for the non-technical audience.
- Background information: The Algorithmic Foundations of Differential Privacy.

Chapter 13

Secure computation

On this page: data-to-code, code-to-data, tools-to-data, algorithm-to-data, cryptography, technique, tool, computing, computation, analysis, analyse, distributed analysis
Date of last review: 2023-04-02

When you use personal data in your research project, you likely also need to analyse those data, often using a script of sorts. In this chapter, we discuss the following scenarios for analysing personal data:



1. "Regular" data analysis ("data-to-code"), where the data are brought to the "script" or analysis software in order to analyse them.
2. "Code-to-data" scenario, where a script or analysis software is run on the data, without moving the data elsewhere.
3. Federated analysis scenario, where a script or analysis software runs on multiple datasets that are in different locations, without moving those datasets elsewhere.

Additionally, we discuss relatively new cryptographic techniques that can be used in securing the analysis of personal or otherwise sensitive data.

13.0.0.1 Which scenario should I choose?

Which scenario is suitable to apply in your project depends on, among others:

- Your dataset: does it contain personal data? How large is the dataset? Do you know the data structure and analysis method beforehand?
- Which computing facility is most suitable:
    - Local (e.g., laptop), on campus (e.g., cluster at Geosciences), from a national trusted party (e.g., SURF), or external (e.g., Amazon, Microsoft)?
    - Located in the Netherlands, Europe or in a non-EEA country?
    - Small or large amount of computing power (CPUs/codes/threads or GPUs, memory size, disk space, etc.)?
- Which software you need to run on the data using the computer power, e.g., R, Python, SPSS, or any other scripting language.
    - Does the software require root user access to install and/or configure?
    - Does the software require paid licenses (e.g., MATLAB)?
    - Can the software be installed in advance, or does it need to be updated during analyses (e.g., with additional packages from a repository)?
- Whether and with whom you are collaborating on your project.

13.0.0.2 Tools and support

We have created an overview of secure computing software and services in this GitHub repository. Keep in mind that this is by no means a complete list!

If you work at Utrecht University, you can ask the Research engineering team for help with choosing a suitable computing solution. If you have already chosen a solution, but are not sure whether it is safe to use, you can contact Information Security or your privacy officer for help.

13.1 “Regular” data analysis: data-to-code

On this page: analysis, data-to-code, data-to-script, transfer, sharing
Date of last review: 2023-04-02

In this scenario, you transfer the data to a computing facility, and run an analysis (script) on the data. In the most basic variant, this computing facility consists of your work computer or faculty computing cluster, where you do not transfer the data outside of your organisation for the analysis. In other cases, data need to be transferred to a computing facility outside your organisation, such as high-performance clusters from SURF, Microsoft, Amazon, etc.

13.1.1 When to use

If you have a relatively small dataset, the "data-to-code" scenario is the most common and flexible scenario:

- It allows you to choose a computing facility that is best suited to your situation.

- It allows you to interactively read, analyse, export and transport the data you want.

Disadvantages of this approach can be:

- When transferring the data to a computing facility, often new copies of the data are created, which can make it more difficult to keep track of different versions of the data.
- Transferring data always comes with additional risks of a data breach. Besides protection during data storage, it is therefore crucial to also protect the data during the transfer to the computing facility, and when used at the computing facility itself.

- The way the data are transferred to the computing facility is not always as straightforward, especially if you have a large dataset.

13.1.2 Implications for research

In this scenario, you need to make sure that:

- You apply data minimisation, access control, and, if applicable, pseudonymisation and other protective measures to limit the amount of personal data that is transferred to the computing facility.
- The data are also protected during the transfer to the computing facility (e.g., your work laptop or an external solution), for example through encryption.

Additionally, if the computing facility is provided by an external processor (e.g., SURF, Amazon):

- A data processing agreement with the provider of the computing facility is needed. If there is none, you cannot use the computing facility to analyse personal data.
- The computing facility should be suitable (secure enough) for the sensitivity level of your (personal) data. For example, if your data are "critical" in terms of confidentiality, the computing facility should also have that "critical" classification.

13.1.3 Examples

- You use your faculty's high performance cluster to analyse a dataset that you collected at your organisation.
- You use the High Performance Computing platform from SURF to analyse a large dataset that you collected at your organisation. In this case, a data processing agreement between your organisation and SURF is needed to make sure that your organisation remains in control of the personal data at SURF's servers.
- You use Amazon Web Services (AWS) to analyse a large dataset that you collected at your organisation. In this case, a data processing agreement between your organisation and AWS is even more important, because Amazon has servers that are located outside of the European Economic Area.

13.2 Code-to-data (one data provider)

On this page: code-to-data, script-to-data, algorithm-to-data, tools-to-data, SANE, digital research environment, secure research environment, virtual research environment, access control
Date of last review: 2023-04-02

In this scenario, an analysis is run on data without transferring the data outside of the organisation In many cases, only the results of the analysis can be exported, and not the data.

We distinguish the following versions of this scenario:

'Tinker' version: interaction with the data

In the Tinker version, users can log in to the computing facility and directly interact with the data, but there may be technical limitations on the import and export of the data. Procedural limitations should be posed through agreements with the user. This version can be implemented in multiple ways, such as:

Accessing and analysing locally stored data on premises. An example is analysing highly sensitive data in a dedicated room without an internet connection.

Accessing locally stored data through remote desktop. This usually does not impose technical limitations on what can be done with the data.

Virtual Research Environments (VREs) are temporary facilities where you can interactively perform computations on data in the cloud. In this case, it is sometimes possible to impose technical limitations on what can be done with the data (in which case these are called “Trusted Research Environments”). Examples of VREs are SURF Research Cloud and anDREa.

‘Blind’ version: remote execution

In the Blind version, users do not have access to the data at all, and only receive the results of an analysis, after reviews by the data owner(s) to ensure that the results do not contain sensitive details. In this case, a synthetic dataset can be provided to write and test the analysis script on, before it is run on the real dataset. This “blind” version could be run in a dedicated environment where researchers can upload their script, but can also be implemented manually, for example when a researcher sends a script by email to be run on a dataset, and receives the results back via email as well (i.e., this is possible when neither the script nor the results contain any sensitive details).

At the moment, both the Tinker and Blind versions of this code-to-data scenario are being developed as virtual research environments in the Secure ANalysis Environment project (SANE).

13.2.1 When to use

Reasons to use this scenario include:

- You want to retain control over the data, e.g., to prevent any unnecessary copies from being made (data sovereignty).
- You do not want, or are not allowed to transfer the data, because they contain personal data or intellectual property.
- The dataset is too large to transfer.
- In the ‘Blind’ version: You want to be sure that the analysis results do not contain any sensitive details.

13.2.2 Implications for research

Compared to the “data-to-code” scenario, the code-to-data approach offers more control over the data, but often requires more, sometimes manual, work, such as:

- Checking the credentials of a user: can they be trusted? An agreement with the user may be desirable or even required. In SURF Research Cloud, credentials can be checked using SURF Research Access Management.

- Preparing a protected computing environment that a user can use.
- In the 'Blind' version:
  - Creating a synthetic dataset.
  - Reviewing the output of the script for sensitive elements. This requires the right expertise.
  - Reviewing whether the code that is run on the data is privacy-preserving. This also requires the right expertise.

It is essential to have a well-described workflow to use this scenario, to ensure confidentiality of the personal data. Additionally, dedicated personnel may make the process easier and consistent.

13.2.3 Examples

- A research team needs to process a dataset containing health data to determine the number of Covid-19 patients at a certain hospital. The hospital providing this dataset does not allow transferring the dataset, but they do allow to run scripts on the dataset. To make that possible, the hospital provides a computing facility, owned by the hospital, to run scripts from research teams. In addition, for each result, the hospital staff inspects if it contains personal data, and if not, it will be passed onto the research team. Since a result like "100 patients at this hospital have had Covid-19 in 2021" does not contain personal data, it can be safely passed to the research team.
- In the data donation approach, the software PORT can be run on data subjects' locally stored data, and only the results of that analysis can be shared with the researcher if allowed by the data subject. Note however that the sensitivity of the results fully depend on the analysis that was run.

13.3 Federated analysis

On this page: federated analysis, federated learning, machine learning, distributed analysis, distributed learning, collaboration, harmonisation
Date of last review: 2023-04-02

Federated analysis is an extension to the code-to-data scenario, where the data of interest are owned by multiple organisations. In this scenario, the data remain with multiple data providers, and the script "travels" across those data sources, combining the results in a central location, and only sharing the results of the analysis. If necessary, there are techniques to hide intermediate results (which could also reveal sensitive information). If the script in question is a machine learning model, then this technique is called "federated (machine) learning". You can learn more about federated analysis in this article.

13.3.1 When to use

Federated analysis is useful when there are multiple data providers who do not allow transferring their data outside of the organisation, or whose data are simply too large to share.

13.3.2 Implications for research

- A prerequisite for analysing data in this way is often that the data at the different providers are similarly structured and use similar terminology (e.g., making sure that every party uses "male", "female", and "other" as levels for the variable Gender, instead of "girl" and "boy", or 0 and 1).
- Federated analysis works best for "horizontally partitioned" datasets, where different organisations have the same (types of) information, but from different people. It is not well-suited for "vertically partitioned" datasets, where the different organisations have different (types of) information on the same people and thus want to link those different data sources.
- Setting up the infrastructure for federated analysis is challenging and can take a large amount of time (software installation, access rights, linking datasets, etc.). It is wise to first investigate whether this option is indeed the most suitable for your project.

13.3.3 Examples

- A research team needs access to various datasets containing health data to determine which factors contribute to health of Covid-19 patients at various hospitals. Each dataset contains health data from patients of the hospital where they are treated. Since each dataset contains sensitive personal data, it is not desirable to store these datasets in a central location to combine them. To be able to answer the research question, one needs to access each dataset separately and combine the results of each dataset. To make this possible, each hospital provides a computing facility. The research team submits their script to each of the computing facilities, where it is run on the local dataset. After a check by each hospital's staff that the results do not contain any sensitive details, the results of the individual computations are combined centrally into one result. In the example, the result of the calculation at each hospital is a prediction model for Covid-19 patients, and the individual models are combined to create a more reliable prediction model.
- Several university medical centres use the Personal Health Train from Health-RI, which relies on the vantage6 software.
- DataSHIELD is an infrastructure and a series of R packages that allows to co-analyse data hosted at different organisations. It requires harmonising

the data at the different organisations and setting up the DataSHIELD infrastructure.

13.4 Cryptographic techniques

On this page: encryption, cryptography, security, collaboration, confidential computing, mpc, homomorphic encryption
Date of last review: 2023-04-02

Besides the scenarios described previously, there are also multiple cryptographic techniques that can be applied to protect sensitive data in the analysis phase. Here, we discuss secure multiparty computation, confidential computing, and homomorphic encryption.

Although there is some overlap in functionality and purpose between these three techniques, they are generally still considered to be distinct and can be combined to enhance security.

These cryptographic techniques are relatively new and are not available as distinct services (yet) for direct application in research. They are for now listed here for information purposes.

13.4.1 Secure multiparty computation

Secure multiparty computation (also referred to as "MPC") is a set of cryptographic techniques that allows multiple parties to jointly perform analyses on distributed datasets, as if they had a shared database, and without revealing the underlying data to each other. Among those techniques are secure set intersection (securely investigating which elements multiple databases have in common), homomorphic encryption (see below), and others.

13.4.1.1 When to use

The benefits of MPC are that no raw data are shared between the parties, computations are guaranteed to perform correctly, and there is a degree of control on who receives the result of the computation (i.e., the results are not necessarily combined in a central location). MPC is therefore a good way of implementing Privacy by Design into your project when you work with personal data.

Contrary to federated analysis, MPC is suitable for linking "vertically partitioned" datasets, i.e., when different organisations have different (types of) information on the same people and thus want to link those different data sources.

13.4.1.2 Implications for research

- The computation in MPC is really joint: you need to have agreed on a specific analysis to be performed and what you will reveal as result of the computation.
- There is no one-size-fits-all MPC solution: different use cases ask for different implementations of MPC.
- Additional computational resources are required to generate random secrets and distribute data over the multiple parties.

13.4.1.3 Example

- MPC was used by a medical insurance company and hospital to determine the effectiveness of a personal lifestyle app for diabetes. In this example, it was possible to calculate average medical cost for different patient groups, based on whether they used the app or not, without revealing patient information between the insurance company and the hospital.
- You can find a simplified example on jointly calculating average income here.

You can find more information about secure multiparty computation on https://securecomputation.org/, in this report, and on the website of TNO.

13.4.2 Confidential computing

Confidential computing is a technique that protects data in use through a (hardware-based) Trusted Execution Environment (TEE). This environment makes sure that data within it are kept confidential (data confidentiality) and that both the data and the code running in the TEE cannot be modified or deleted (data and code integrity). The TEE uses embedded encryption keys and makes sure that the analysis stops running when malware or unauthorised access is detected. Moreover, data and code are even invisible to the operating system, cloud provider and any virtual machines.

There are many possible applications of this technique, for example:

- You want to protect against unauthorised access during the analysis of sensitive data.
- You want to analyse sensitive data, and it is necessary to use an untrusted cloud platform or infrastructure.
- You want to prevent the analysis script from leaking or manipulating data.

It is important that confidential computing is used together with encryption of data at rest and in transit, with restricted access to the decryption keys. It also requires the TEE to be trustworthy (attestation), which is an active field of study. You can read more on the website of the Confidential Computing Consortium.

13.4.3 (Fully) homomorphic encryption

Where "regular" encryption focuses on data at rest (e.g., in storage) or data in transit (e.g., when transferring data), homomorphic encryption allows analyses to be performed on encrypted data ("data in use"). During the analysis, both the data and the computation result remain encrypted, unless they are decrypted by the decryption key owner. This technique can be applied both in confidential computing and in secure multiparty computation.

There are multiple types of homomorphic encryption: partial, somewhat partial and fully homomorphic encryption. The latter is the most promising solution, as it allows an infinite number of additions and multiplications to be performed on the encrypted data.

Currently, the practical use of homomorphic is limited, because it can require a lot of computational resources to use it, causing it to be relatively slow. New implementations are however being developed, see this website for a list of available implementations. Another limitation is that there is no interaction with the data during the analysis, and so you cannot check whether the analysis was successful. To solve this, you could use a synthetic dataset to develop and test your algorithms first.

Chapter 14

Other techniques

In addition to the techniques discussed in the previous chapters, other techniques exist that can help protect data subjects' privacy. Here, we discuss the following techniques:

- Encryption, for example of files, folder or entire drives.
- Synthetic data that does not contain data from real individuals, but can mimic statistical properties of the original dataset.
- Data donation, in which citizens are asked to donate digital trace data for scientific research purposes.

14.1 Encryption

On this page: encryption, cryptography, cryptographic technique, secure storage, encryption software
Date of last review: 2023-05-15

Encryption is a technique to convert digital information into a code or cipher, which can only be read by someone who has the key to decipher or decrypt it. It can be applied to many digital objects, such as text strings, files, folders or entire storage drives. The format of the decryption key can also vary between a password, a randomly generated code, or a file.

For personal data, encryption can be seen as a pseudonymisation technique, where the encrypted data are pseudonymised and the encryption key is the additional information needed to identify individuals. In research, encryption is often applied for data "at rest", that is, data that are stored and not actively used. However, data can also be encrypted *in transit* (i.e., during transfers) or even *in use* (i.e., performing computations on encrypted data).

14.1.1 Types of encryption

There are several types of encryption. How they work can get complicated very quickly, but here is a general overview of the different types:

1. **Symmetric encryption**: the same key is used for both encryption and decryption. This is a relatively quick way to encrypt data and is most often used for research data. Because only one key is needed to leak the data, a hard-to-guess key, secure storage and secure transfer of the key is crucial. Example algorithms that use symmetric encryption are AES, (3)DES, Blowfish, and IDEA.
2. **Asymmetric encryption** (public-key): two different keys are used for encryption and decryption: a public key is used for encryption and can be shared with anyone, and a private key is used for decryption, which must be kept secret. This is also known as end-to-end encryption and is used in many messaging platforms to prevent service providers from decrypting private messages. Example algorithms are RSA and elliptic curve cryptography.
3. **Hybrid cyphers**: hybrid cyphers combine the speed of symmetric, and the security of asymmetric encryption. Typically, a symmetric algorithm is used to encrypt the data, and an asymmetric algorithm to encrypt the symmetric key. This type of encryption is commonly used in secure communications, such as email and virtual private networks (VPNs).

In general, the more “bits” that are used by an encryption algorithm, the larger the number of possible keys is, and thus the harder it is to guess the correct key using a brute-force attack.

14.1.2 When to use

Encryption can be applied on different levels. In research, data are encrypted usually either on a drive-level or on file/folder level:

- **Full-drive encryption** (“volume” encryption) makes sure that data on storage drives or devices are not readable if someone gains unauthorised access to the device or drive. This is generally recommended to always apply to devices that contain research data, but particularly when:
  - you want to protect data on your personal laptop (encrypt the entire laptop or specific hard drives).
  - you collect data on portable devices like USB sticks and audio recorders.
- Encryption of **individual files or folders** (“container” encryption) can be used when you need to protect individual files or folders. Use it when:

  - you cannot physically separate different types of personal data on different storage locations and need to make sure that a limited number of people can access the encrypted data.
  - you have to store personal data on a non-encrypted drive that multiple people have access to, which they do not need.
  - you need to send personal data to a collaborator, for example via the cloud or via a file sender.

14.1.3 Implications for research

- Encryption only guarantees protection while the data are encrypted, which is usually during storage or in transit. For example, encryption is generally not a suitable safeguard to protect data during data analysis, because usually data need to be decrypted in order to be read by analysis software. This implies that when you need to decrypt the data, other safeguards must be in place to protect the data, such as controlled access and a secure workspace.
- Responsible key management is crucial. In principle, the data cannot be accessed without a decryption key. Although some encryption software offers the possibility to create recovery keys, there is still someone who needs to manage the key(s), as long as the data are encrypted.
- When you use file/folder encryption, please note that in some cases, folder and/or file names are not encrypted. If these names contain sensitive information, consider renaming the folders/files or putting them in a zipped folder with a non-sensitive name and then encrypting the zipped folder.
- Just because you encrypt data with a state-of-the-art encryption algorithm now, that does not mean the data are necessarily protected in the future as well. Encryption algorithms can become unsafe due to, for example, new technological developments or bugs in the encryption algorithm that make it more vulnerable to attacks. Therefore, if you store encrypted files for a long period of time, you need to regularly re-assess whether the algorithm is still secure enough and if needed, re-encrypt the data.

14.1.4 Tools and resources

- We have created an overview of commonly used encryption tools in our GitHub repository. **Note**: many institutes have institution-wide encryption software available. Please consult with your information security officer to determine which tool you can best use for your situation.
- Ghent University has created a guidance on several common encryption tools in different scenarios.
- In this introductory book, you can read further if you want to know more about cryptography.

14.2 Synthetic Data

On this page: synthetic data, fake data, artificial data, dummy data, fictitious data, reproducibility, reproduce, open science workflow
Date of last review: 2023-05-15

Synthetic data generation is the process of creating artificial data, which can be used in place of the real, possibly personal, data. Instead of simply adjusting an existing dataset to make it less identifiable, a completely new dataset is generated, with fictitious individuals. When creating synthetic data, sensitive values (which can be some values, or the entire dataset) in the data are replaced by values that are generated from a statistical model. The intuition behind this is to mimic the idea of drawing samples from a population. Not with actual people, but with fictitious individuals that "look like" the people from the population. Synthetic data can be created in multiple ways, such as based on rules or using a trained machine learning model, and for different purposes, such as for privacy protection purposes, but also for data enrichment or software testing.

14.2.1 When to use

Synthetic data can be generated for a variety of reasons, for example:

- As an intermediate step in sharing (personal) data, before others gain access to (part of) the real dataset. This can for example be useful when data recipients still have to determine which variables they need from the real dataset, or how many observations. Or when the data request procedure can take up a large amount of time and recipients already want to explore the (synthetic) data.
- To develop code without requiring access to real (personal) data. In this case, the synthetic data usually does not need to mirror the data statistically, but just in terms of the structure (e.g., only with the same column names and data types). Synthetic data that only resembles the real dataset in terms of its structure is also sometimes called "dummy data".
- To adhere to an open science workflow, to evaluate and reproduce analyses. If you share both a synthetic dataset and your code, others can easily evaluate your code with an actual dataset, see the results of the code and test its reproducibility.
- In teaching, to prevent having to share (personal) data with students.

14.2.2 Implications for research

- Although synthetic data are artificial, privacy risks may still remain. You can think of it like this: if the generating model is too good, you could reproduce the original data exactly, doing no better in protecting anyone's privacy. On the other end, you can create a synthetic dataset that sets every value to "0", which protects privacy very well, but is useless in terms of its quality. Usually, the result is somewhere in between: the synthetic data is less informative than the real data, at the expense of leaking some information about the original sample (e.g., descriptive statistics, relationships between variables, plausible values in the data). Hence, the synthetic data lies on a **privacy-utility spectrum**. Whether the synthetic dataset still contains personal data will need to be considered on a case-by-case basis and will differ depending on the method used to create the synthetic dataset. For more detailed information on this concept of the privacy-utility spectrum, see the UK Office for National Statistics.

- Synthetic data is sometimes used in conjunction with Differential Privacy, which can help to numerically set the level of privacy/utility. Unfortunately, it is not straightforward to determine the disclosure risk directly from the synthetic data, and it is better to fix the privacy leakage in the process of generating the data.
- The quality of the synthetic dataset is highly dependent on the input from which it is created. In general, a larger number of rows in the dataset (individuals) results in better quality synthetic data as there is more variability in the dataset. Moreover, datasets that contain many outliers typically result in lower-quality synthetic data, because they can have a large influence on the statistical properties of the dataset (e.g., the mean) from which the synthetic dataset is created. This in turn can lead to a distorted or unrealistic distribution in the synthetic dataset.

14.2.3 Tools and resources

- We have created an overview of tools to create synthetic data in our tool repository on GitHub.
- Here's a comprehensive tutorial for creating synthetic data that includes information on the Python package MetaSynth (workshop) and the R package Synthpop (how-to).
- Read further about synthetic data in "Synthetic data - What, Why and How?"
- The United Nations has created this Starter Guide on synthetic data, which includes what to consider when creating synthetic data.
- If you want to know more about the methodology behind synthetic data, we can recommend the book "Synthetic Datasets for Statistical Disclosure Control" (Drechsler, 2011).

14.3 Data donation

On this page: data donation, digital trace data, local processing, interactive consent
Date of last review: 2023-05-15

People leave all kinds of digital traces, for example on social media, smartphones, search engines, email, banks, energy providers, and online shops. Article 15 of the GDPR mandates that individuals have the right to request access to a copy of their personal data collected and stored on those digital platforms. The owners of those platforms will then make that digital trace data available through individual "Data Download Packages" (DDPs). These DDPs can contain a lot of personal information that may be too sensitive to share, but can also be of high value for scientific research purposes.

The data donation approach as developed by Boeschoten et al. (2020) allows researchers to automatically analyse the digital traces found in DDPs, while preserving the privacy of research participants. The workflow is as follows:

1. Data subjects are recruited as respondents like in a regular research project.
2. The researcher determines which DDPs are relevant for the research question and writes a script to extract the relevant information.
3. Each data subject requests their DDPs with the selected providers and stores these locally on their own device.
4. Stored DDPs are then locally processed with the software PORT to extract relevant research variables. PORT executes the script provided by the researcher and locally extracts the data from the DDP. PORT uses Pyodide technology to run in its own secure environment which is completely separated from the device. This environment is destroyed as soon as the browser page is closed.
5. The data subject inspects the information resulting from the analysis and is asked to provide informed consent to share it with the researcher.
6. If the data subject consents, the derived information is encrypted and sent to the researcher for further analysis.

14.3.1 When to use

The data donation approach can be used:

- As an alternative or in addition to surveys to study human behaviour.
- To analyse data that are too sensitive to transfer in raw form.
- To allow data subjects a large degree of control over their (personal) data.
- To access data that are representative of a population of interest, in contrast to, for example, data retrieved from APIs, which often pertain a non-random subset of a platform's user group.

- As a user-centric approach that is independent from platforms or data controllers: private companies cannot suddenly withdraw from a collaboration or restrict access to a dataset, because the data were not obtained directly through them. It is important, however, to review the Terms of Service of the platforms you use, to review if there are restrictions on data usage for scientific research.

14.3.2 Implications for research

- DDPs may be large and contain different types of information. For both the analysis (writing the script) and the informed consent (informing data subjects specifically), it is important that you know which specific data are of interest.
- The structure of DDPs varies by provider and by person, making it difficult to set up analysis scripts generically. Moreover, DDPs change over time. Analysis scripts should regularly be checked and updated (Boeschoten et al., 2021).
- Analysis scripts are usually developed based on sample data. However, due to the sensitive content, DDP sample data are difficult to obtain. As sample data, you could use your own DDP, synthetic data (example), or already available open data (example).
- It is important to make sure that data subjects understand what they are consenting to when presenting the results that will be shared with you (step 5). Do they understand the risks involved (if any)? We recommend talking to a privacy officer and/or testing this among data subjects before you start your Data Donation project.

14.3.3 Examples and resources

- Several projects have made use of the data donation approach, such as one using Google semantic location history data, and one with Whatsapp data.
- A more elaborate data donation platform is being developed in the PDI-SSH-funded Digital Data Donation Infrastructure (D3I) project.
- Read further about the framework, the proof of concept of the PORT software, a comparison with a browser plug-in approach, and promises and pitfalls of the approach for social media data.
- Or read more about how Data Download Packages can be de-identified.

Chapter 15

Tools & Services

There are many tools that you can use to work with personal data, such as tools to collect, store, and analyse personal data, but also to deidentify, encrypt and synthetise datasets. In this chapter, we provide resources to identify the tool you are looking for.

In short:

- If you work at Utrecht University, you can use https://tools.uu.nl to find UU-approved tools.
- In this GitHub repository, you can find tools and packages to deidentify, encrypt, synthetise and otherwise work with personal data.
- If you are in doubt whether you can use a specific tool, please contact your privacy officer, data manager or information security for help.

Are you developing a website or application yourself that uses user data? Check out the CNIL GDPR Guide for Developers for step-by-step guidance on how to develop your software in compliance with the GDPR.

15.1 Utrecht University tool finders

On this page: storage, storing, collection, repository, data archive, sharing, tools, services
Date of last review: 2023-02-17

When you are using a tool that processes personal data, that tool should do so in compliance with the GDPR. If you work at Utrecht University (UU), you can use https://tools.uu.nl to find:

- tools that are safe to use in the Tooladvisor. These include tools for data collection, file sharing, audio transcription, and more. Most of the tools listed in the Tooladvisor are safe to use either because no (personal) data are being used by the tool, data are processed at UU premises, or because of a Data Processing Agreement between UU and the supplier of the tool, in which the supplier agreed to sufficiently protect the data entered into their tool.
- all storage facilities provided by UU.
- a selection of possible data repositories to publish (meta)data in.

Additionally, you can find available software via this intranet page

15.2 Tools to deidentify, synthetise and work safely with personal data

On this page: anonymisation, pseudonymisation, de-identification, synthetic data, encryption, secure computing, computation
Date of last review: 2023-02-17

We are creating an overview of potential privacy-related tools for deidentifying data, creating synthetic data, and analysing data in secure environments in a GitHub repository.

To the overview

Please feel free to open an Issue or a Pull Request in this repository if you wish to adjust the existing content or add new content.

15.3 Requirements for a third-party tool

On this page: custom tool, provider, agreement, third-party tool, service
Date of last review: 2023-02-17

If your tool of choice is not listed in https://tools.uu.nl, but it does process personal data, please contact the IT servicedesk. They will help you assess whether a tool is safe to use.

If a tool is processing personal data, the following two aspects are important to consider:

15.3.1 1. Who is processing the personal data: arrange an agreement

When you use a third-party tool that processes personal data, the data are not under your (full) control. In this case, you must ensure the GDPR compliance of the tool provider using (art.46):

- A Data processing agreement - when the provider processes (e.g., stores, analyses, collects) personal data within the European Economic Area (EEA) or a country with an adequate level of data protection.
- Standard contractual clauses (SCCs) - when personal data are processed by a supplier outside of the EEA without an adequate level of data protection. These make sure the provider will use sufficient measures to protect the personal data and enable data subjects to exercise their rights.
- Explicit consent of data subjects who have been informed on the risks involved - in the absence of an agreement. Please contact your privacy officer if you are considering this option.

You can assume agreements are in place for the tools recommended by UU. If there is no agreement in place between UU and the tool provider, using this tool is **not allowed**, even if the provider is located within the EEA, has an adequate level of data protection, or has high security standards. The only exception is when data are always end-to-end encrypted, because then the tool provider cannot learn anything from the data.

15.3.2 2. Security level

The tool provider should employ good security practices, such as regular backups in distinct geographical areas (preferably in replication rather than on tape), regular integrity checks, encryption at rest, multi-factor authentication, etc. Most of these aspects will likely be covered in the agreement, and sometimes a data classification will need to be performed. Information security can help you determine all necessary security requirements.

Use Cases

Chapter 16

Data minimisation in a survey

On this page: minimise, limit, remove, questionnaire, survey Date of last review: 2022-08-22

For a course, a teacher at the faculty of Veterinary Medicine collected data on the health of pets and the pets' owners. The initial purpose of the survey was to create simply datasets for students to learn about statistics. However, besides for the course, the teacher also wanted to use the collected data for research purposes and share the data with others. In order to do so, the teacher created a new version of the survey that asked for less identifiable information and could be more easily anonymised. Additionally, the new version of the survey informed participants about the legal basis used to process their personal data.

Here, you can find the survey before and after data minimisation:

Before minimisation After minimisation

Note that the new version of the survey:

- minimises the amount of personal data collected:
    - Student number and pet names are not asked in the new version of the survey.
    - Instead of Age, the new version asks the Age category of the owner/caretaker.
    - The survey includes questions on Weight and Height. For data publication, they are used to calculate the Body Mass Index (BMI) and deleted after this calculation.
- contains information about the legal basis used to be able to use (legitimate interest) and publish (consent) the data for purposes other than education.

Chapter 17

Data pseudonymisation

On this page: pseudonymous, de-identification, replacement, open science, reuse
Date of last review: 2023-03-30

YOUth (Youth of Utrecht) is a longitudinal child cohort study that collects data about the behavioural and cognitive development of children in the Utrecht area. The study follows about 4000 children and their parents in two cohorts. One from birth until around the age of six, one from around 9-years-old until adolescence. YOUth collects a wide variety of data types, ranging from questionnaires to biological samples. Because of the large amount of data and the sensitive nature of the data and the participants (minors), the data can be considered as very sensitive, and thus should be pseudonymised where possible.

17.0.1 General steps

YOUth is committed to sharing their data for reuse, and thus the datasets that they share should contain as little personal information as possible. For that purpose, the YOUth data manager implements a number of measures:

- All data are pseudonymised as much as possible (see below).
- Every dataset that is shared for reuse is first checked for identifiable information. Special category information is taken out of the datasets as much as possible, and no unnecessary information such as date of birth is shared.
- Using the tool AnonymoUUs, participant pseudonyms are replaced with artificial pseudonyms, and all dates with a fake date, each time a new set of data is prepared for sharing. This limits the ability of external researchers to link multiple requested datasets together and thus to form a more complete image of each participant. It also prevents singling out participants based on the day they visited the research centre.

17.0.2 Pseudonymisation per data type

Below is an overview of the data types and pseudonymisation measures taken by the YOUth data manager. Besides these pseudonymisation measures, YOUth has implemented a data request procedure which delineates the conditions under which researchers can access the data, and the steps they have to take to request access.

Questionnaire data (tabular)

Children and their parents/caretakers (sometimes their teacher) fill out several questionnaires about, among others, their mental and physical development, living conditions, and social environment.

Pseudonymisation measures:

A script removes unnecessary (special category) personal data from the shared dataset where possible, such as religion, ethnicity and open text responses.

If a researcher needs demographic information only to describe the sample, the data manager shares a frequency table of the requested information, for example for ethnicity and socio-economic status, instead of sharing the raw responses.

The Anonymouus tool replaces the pseudonym and date in the questionnaire data and file names.

In the future, the data manager would like to share only scale scores, instead of responses to individual questions in standardised questionnaires.

Computer tasks (tabular)

On a computer, children play various games to measure cognitive and motoric development of the child. In most games, the response times, choices and scores are recorded. To pseudonymise the data, the AnonymoUUs tool replaces the pseudonym and dates in the task data and filenames and in some cases even the name of the participant.

Logbook- and experiment book data (tabular)

Notes about data collection (data quality, task-order, if experiment started etc.) are made in logbooks by means of a data capturing tool. In that same tool, YOUth also collects research data about body measures (length, weight and head circumference) and intelligence (WISC and WPPSI) To pseudonymise those data, the AnonymoUUs tool replaces the pseudonym and date in the filenames and data.

Video tasks (video recording)

During two tasks (the Hand game and the Delay of gratification task), children are video- and audiotaped to be able to analyse their behaviour. Parents may also be visible in the background, as well as a research assistant.

To pseudonymise these data, both the videos from the Hand game and the Delayed gratification task will be coded/scored on the variables of interest (e.g., does the child take the candy out of the bag or not). This way, no actual video recordings will need to be shared with other researchers.

Parent-child interaction (video recording)

Children and their parents are videotaped while they play with each other or discuss specific topics. Because these data are difficult to pseudonymise and could be scored/coded on many different aspects, YOUth provides a special local laboratory space to perform the desired qualitative analysis on these video data.

Magnetic Resonance Imaging (MRI) data (3D image)

MRI data of children are collected to study structural (3D image of the brain, skull, and outer layers of the head) and functional (brain activity) properties of the brain.

To pseudonymise the MRI data, structural MRI scans (DICOM) are defaced using mri_deface (v1.22), resulting in NIfTi files. Additionally, the AnonymoUUs tool replaces the pseudonym in the filenames.

Electro-encephalography (EEG) data (video and text files)

A cap is placed on the child's head with electrodes attached to measure brain activity. The child is placed in front of a monitor and views various on-screen stimuli (incl. faces, objects, sounds, music, toys). A video is also made to check whether the child watches the screen. For the moment, the videos will not be shared with external researchers. In the EEG data itself, the AnonymoUUs tool replaces the pseudonym and date.

Eye tracking data (text files)

Children are placed in front of a screen and view various stimuli (incl. faces, objects, sounds, music, toys), with or without an assignment. Eye movements and focus points are recorded using an eyetracker. To pseudonymise these data, the AnonymoUUs tool replaces the pseudonym and date in the eyetracking data and the filenames.

Ultrasound images (3D echos)

During the mothers' pregnancy, 3D ultrasound images are made of foetuses to follow overall and brain size development. To pseudonymise these data, the ultrasound images (DICOM) will be converted to nifti (.nii) format, which does not contain header information. Additionally, the AnonymoUUs tool replaces the pseudonym and date in the filenames and in the SQL database that comes with the measurement.

Biological materials

At various moments during the study, (chord) blood, hair, saliva, and buccal swabs are taken from the child and sometimes their parent(s). The samples cannot be pseudonymised, because they are physical samples. Instead, a procedure is in place to have biological samples analysed at preferred partners, without having to share the physical samples with researchers.

Chapter 18

Publishing metadata

On this page: FAIR data, metadata, documentation, publication, reuse
Date of last review: 2022-08-22

In 2020, the Open Science Programme of Utrecht University sent out the first Open Science monitor. The aim was to gain insights into the awareness, attitudes, practices, opportunities and barriers of employees of Utrecht University and Utrecht University Medical Center regarding several Open Science practices. As the dataset contained a lot of demographic information (e.g., gender, age, nationality, position, type of contract, etc.), and all of those variables combined could lead to identification, it could not be shared publicly. For this particular dataset, full anonymisation was not desirable, as that would greatly decrease its scientific value. Therefore, the Open Science Programme chose to publish only the metadata and documentation, without sharing the data, in order to protect participants' data while still complying with the FAIR principles.

Here's the strategy they took:

- They published the dataset under restricted access on Yoda, so that the dataset was at least Findable and Accessible.
- They shared other relevant documentation publicly:
  - The preregistration.
  - The questionnaire itself, including the information provided to participants.
  - The final report written about the dataset.

Note that in the metadata of all these publications, cross-references to the other publications are included to allow for maximum findability of the project's outputs.

Chapter 19

Reusing education data for research

On this page: further processing, secondary use, reuse, student data, education, legal basis, access control
Date of last review: 2022-11-18

A research group at the Science faculty wanted to investigate the effects of the Covid-19 pandemic on students' motivation and study success in a specific course. To do so, they wanted to analyse:

- Students' evaluations of the course from both before and during the pandemic.
- Students' test and final grades in the course from both before and during the pandemic.

The primary researchers already had access to the data for their educational activities, and so they wanted to use the data for research purposes. They went to their faculty privacy officer to find out how they could reuse these data in a responsible way.

The following privacy issues are relevant in this use case:

- **The raw data were identifiable** The student grades were linked to names, and both the grades and the evaluations were linked to student IDs. Moreover, the evaluations could potentially contain names of teachers and other personal information, as they consisted of partly open-ended questions. To decrease identifiability, the principal investigator and a second examiner, who already had access to the students' data, first removed or replaced all names with pseudonyms (both names of student and teachers), and went through the open-ended questions to remove potentially directly identifiable information. Only after deidentification were the data shared with research assistants who performed the main data and content analyses.

- **Data subjects' rights** Most students had already finished the course, and were not informed about the use of their evaluations and grades for this research project. The researchers argued that the majority of the students could not be traced anymore to provide this information or to enable them to exercise their data subjects' rights (art. 14(5)(b)). Moreover, in case a student did want to exercise their rights, it would prove difficult to retrieve the correct data, as the data were deidentified as soon as possible.
- **Legal basis** Students did not provide explicit consent to process their grades and evaluations for this research project. Moreover, if they had provided consent, it could be argued that the consent was not freely given, as the primary researchers were also involved as teachers, and therefore there was a hierarchical relationship between the students and the teachers. For these reasons, consent was not a suitable legal basis in this case. Instead, the researchers relied on:

  – **Public interest**: processing students' data for the course itself is a public task, namely that of providing education. It was the legal basis for the initial data collection.

  – **Further processing for scientific research purposes**: processing data to answer the research question can be considered as secondary use of the students' personal data. The GDPR does not consider secondary use of personal data for scientific research purposes incompatible with the original purpose (i.e., the original purpose being to provide education and improving the course, art. 5(1)(b)). Thus, it was not necessary to rely on a new legal basis for this research project, provided that the data were protected sufficiently: The researchers made sure that the data were well-protected (i.e., minimised, pseudonymised, and access controlled, art. 89).

Resources

Chapter 20

Seeking help at Utrecht University

Date of last review: 2023-05-22

If you work at Utrecht University, there are several ways to look for further support.

20.0.1 Education

Research Data Management Support currently offers:

- Online module Privacy basics for researchers
- In-person workshop Handling Personal Data

Additionally, your own faculty or department may offer workshops surrounding privacy, ethics and/or research integrity (see websites below).

20.0.2 Online information

Besides this Handbook, you can find more information on the following websites:

- RDM Support website
- Intranet pages on privacy, research and information security
- Faculty-specific web pages:
  - Geosciences: RDM and privacy, ethics
  - Science: RDM and privacy, ethics

- Social and Behavioural Sciences: tech support, ethics
- Humanities: RDM and privacy, ethics
- Law, Economics and Governance: ethics
- Veterinary medicine: Research Support Office
- Medicine: ethics, data management-related information can be found on UMCU Connect.

20.0.3 In-person support

The **first point of contact** about privacy is the privacy officer of your faculty.

Besides the privacy officer, you can also ask for help from:

- Your local data steward/data manager (see websites above) or Research Data Management Support.
- Information security.
- In some faculties, the Research Support Office may be of help in drafting agreements.

The glossary consists of frequently used jargon concerning the GDPR and research data. Click on a term to see its definition.

A

Anonymous data

Any data where an individual is irreversibly de-identified, both directly (e.g., through names and email addresses) and indirectly. The latter means that you cannot identify someone:

by combining variables or datasets (e.g., a combination of date of birth, gender and birthplace, or the combination of a dataset with its name-number key)

via inference, i.e., when you can deduce who the data are about (e.g., when "profession" is Dutch prime minister, it is clear who the data is about)

by singling out a single subject, such as through unique data points, e.g., someone who is 210 cm tall is relatively easy to identify)

Anonymous data are no longer personal data and thus not subject to GDPR compliance. In practice, anonymous data may be difficult to attain and care must be given that the data legitimately cannot be traced to an individual in any way. The document Opinion 05/2014 on Anonymisation Techniques explains the criteria that must be met for data to be considered anonymous.

C

Controller

The natural or legal entity that, alone or with others, determines or has an influence on **why** and **how** personal data are processed. On an organisational level, Utrecht University (UU) is the controller of personal data collected by UU researchers and will be held responsible in case of GDPR infringement. On a practical level, however, researchers (e.g., Principal Investigators) often determine why and how data are processed, and are thus fulfilling the role of controller themselves.

Note that it is possible to be a controller without having access to personal data, for example if you assign an external company to execute research for which you determined which data they should collect, among which data subjects, how, and for what purpose.

D

Data subject

A living individual who can be identified directly or indirectly through personal data. In a research setting, this would be the individual whose personal data is being processed (see below for the definition of processing).

E

European Economic Area (EEA)

The member states of the European Union and Iceland, Liechtenstein, and Norway. In total, the EEA now consists of 30 countries. The aim of the EEA is to enable the "free movement of goods, people, services and capital" between countries, and this includes (personal) data (source: Eurostat).

G

General Data Protection Regulation (GDPR)

A European data protection regulation meant to protect the personal data of individuals, and facilitates the free movement of personal data within the European Economic Area (EEA). The Dutch name of the regulation is "Algemene Verordening Gegevensbescherming" (AVG).

H

Hashing

Hashing is a way of replacing one or multiple variables with a string of random characters with a fixed length. It can be used to create a "hashed" pseudonym, or to replace multiple variables with one unique value. It is usually quite difficult to reverse the hashing process, except if an attacker has knowledge about the type of information that was masked through hashing. To prevent reversal, cryptographic hashing techniques add a "salt", i.e., a random number or string, to the hash (the result is called a "digest"). If the "salt" is kept confidential or is removed (similar to a keyfile), it is almost impossible to reverse the hashing process.

L

Legal basis

Any processing of personal data should have a valid legal basis. Without it, you are now allowed to process personal data at all. The GDPR provides 6 legal bases: consent, public interest, legitimate interest, legal obligation, performance of a contract, and vital interest. Consent and public interest are most often used in a research context.

P

Personal data

Any information related to an identified or identifiable (living) natural person. This can include identifiers (name, identification number, location data, online identifier or a combination of identifiers) or factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of the person. Moreover, IP addresses, opinions, tweets, answers to questionnaires, etc. may also be personal data, either by itself or through a combination of one another.

Of note: as soon as you collect data related to a person that is identifiable, you are processing personal data. Additionally, pseudonymised data is still considered personal data. Read more in What are personal data?.

Processing

Any operation performed on personal data. This includes collection, storage, organisation, alteration, analysis, transcription, sharing, publishing, deletion, etc.

Processor

A natural or legal entity that processes personal data on behalf of the controller. For example, when using a cloud transcription service, you often need to send personal data (e.g., an audio recording) to the transcription service for the purpose of your research, which is then fulfilling the role of processor. Other examples of processors are mailhouses used to send emails to data subjects, or Trusted Third Parties who hold the keyfile to link pseudonyms to personal data. When using such a third party, you must have a data processing agreement in place.

Pseudonymous data

Personal data that cannot lead to identification *without additional information*, such as a key file linking pseudonyms to names. This additional information should be kept separately and securely and makes for de-identification that is reversible. Data are sometimes pseudonymised by replacing direct identifiers (e.g., names) with a participant code (e.g., number). However, this may not always suffice, as sometimes it is still possible to identify participants indirectly (e.g., through linkage, inference or singling out). Importantly, pseudonymous data are still personal data and therefore must be handled in accordance with the GDPR.

S

Special categories of personal data

Any information pertaining to the data subject which reveals any of the below categories:

racial or ethnic origin

political opinions

religious or philosophical beliefs

trade union membership

genetic and biometric data when meant to uniquely identify someone

physical or mental health conditions

an individual's sex life or sexual orientation

The processing of these categories of data is **prohibited**, unless one of the exceptions of article 9 applies. For example, an exception applies when:

the data subject has provided explicit consent to process these data for a specific purpose,

the data subject has made the data publicly available themselves,

processing is necessary for scientific research purposes and obtaining consent is impossible or would require an unreasonable amount of effort.

Contact your privacy officer if you wish to process special categories of personal data.

T

Third-country transfer

In legal terms, a transfer exists when personal data controlled by one party are accessible to another, irrespective of whether the data are physically sent to that party. An international/third-country transfer exists when the party that can potentially gain access is based in a country outside the European Economic Area (EEA) which does not have an adequacy decision from the European Commission.

Chapter 21

Resources

For further reading, we prepared a Zotero library with additional resources, some of which are specific to Utrecht University, others more general. Click on the image below to see the most recent version of the reference library online.