Guide for data protection in research

This guide aims to support University of Tartu researchers in complying with the Code of Conduct for Research Integrity. As that inevitably requires some knowledge of data protection principles, the guide deals with the ethical handling and processing of personal data in research.

The guide explains how legislation related to personal data protection affects research work and points out the exceptions. Reading the guide does not require any previous knowledge of data protection: all the relevant concepts and general principles are explained in the first chapter.

The guide is further structured according to data lifecycle: the second chapter deals with the planning of research, the third with data collection and analysis, and the fourth with publication and retention of data. If you cannot find an answer to a specific question in this guide, you are welcome to contact the data protection specialists of the University of Tartu at andmekaitse@ut.ee.

Authors: Marten Juurik, Terje Mäesalu and Tiiu Tarkpea.

Editor: Helika Mäekivi

Contents

1. Basic concepts of data protection

This chapter explains the basic data protection concepts and how they relate to research. It also gives a brief overview of how personal data processing is regulated by law and how it relates to research ethics.

1.1. What is personal data processing?

The aim of protecting personal data (also: 'data protection') is to protect individuals and their privacy. The right to the protection of personal data is laid out in the Treaty on the Functioning of the European Union and the Charter of Fundamental Rights of the European Union as an independent fundamental right, which unambiguously demonstrates its relevance and importance.

For the purposes of this guide, **data protection** is a field of law governing the use of personal data. On the one hand, it includes principles that may seem self-evident: for example, the principle of voluntariness, which means that individuals must not be asked to consent to the processing of their data by threat or coercion; or respecting the person's autonomy, which means that everyone should be able to control the use of their personal data. On the other hand, personal data protection conflicts with several other important goals and interests, such as maintaining public order, doing research or business. There are exceptions, rules and principles for dealing with conflicting interests, but they may not be self-evident.

This guide explains the exceptions that apply to processing personal data in scientific research. It is important to remember that the data protection field is broad, and new issues are constantly emerging with advances in society and technology.

Data protection at the University of Tartu

Short guidelines on various topics of personal data protection are available on the University of Tartu's intranet. The university's wiki pages provide an overview of the main data protection principles by topics (in Estonian). A data protection module for testing and improving one's knowledge has been prepared for university staff. The data protection policy is available on the university's website. Chapter IX of the university's documentary procedure rules explains the university employees' rights and responsibilities when processing personal data. The university's public wiki pages describe what to do in the event of information security and data protection incidents.

1.2. What are the main legal acts governing personal data protection?

Based on Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (or General Data Protection Regulation, GDPR), the general data protection principles have been established for the processing of any kind of personal data, including in research. With the entry into force of the GDPR, many significant changes were made in the legal framework of data protection.

Firstly, the GDPR is directly applicable, i.e., it applies as written and unlike directives, does not have to be transposed into Estonian law or established by a separate act. However, it must be implemented. In addition, around 130 legislative acts have been amended in Estonia to align them with the GDPR.

The GDPR provides general principles agreed upon by the EU member states but does not offer individual solutions for specific situations. That is why numerous guidelines, recommendations and best practices have been compiled to help people understand the provisions of the GDPR. These guidelines also aim to draw attention to respecting the obligation to protect personal data when conducting research.

Secondly, each EU member state can **specify** certain important **exceptions by law**. The processing of personal data in research is one of the areas that each member state regulates by national law. In Estonia, it is regulated by the Personal Data Protection Act, which, for example, specifies the requirement of the ethics committee's approval that is not included in the GDPR. Thus, processing personal data in research is more complex, as both the GDPR and the Personal Data Protection Act must be complied with and understood.

What changed with the entry into force of the GDPR?

During the data protection reform in the European Union, several previously existing principles were clarified, and additional requirements were introduced for data controllers. In the following, some of the most important changes concerning research have been listed.

Changes in terminology: the former *coding* and *decoding* were replaced by *pseudonymisation*; instead of *sensitive personal data*, *special categories of personal data* were introduced in the GDPR. The phrase is also used as "*personal data of special categories*" in the Personal Data Protection Act.

Registration of controllers is no longer required. Previously, there was a system in Estonia where processors of special categories of personal data had to register with the Data Protection Inspectorate. Now, there is no such requirement. There is a general requirement to consult an ethics committee or, in the absence of a relevant committee, the Data Protection Inspectorate, to ensure the due processing of special categories of personal data.

New obligations of the controller. As a data controller, the University of Tartu has received several additional responsibilities: to publish its data protection policy, prepare an overview of the processing of personal data, appoint a data protection specialist, prepare data protection impact assessments, consult supervisory authorities and report breaches.

The processing of public personal data is limited. Previously, an exception was established under the Personal Data Protection Act, always allowing further processing of lawfully disclosed personal data.[1] Such exception no longer exists, which means that a legal basis is also necessary for processing previously disclosed personal data; otherwise, the processing is unlawful.

Fines were established. The Data Protection Inspectorate was given additional authority to impose fines for data protection breaches, up to 4% of total annual global turnover for companies and up to 20 million euros for other persons in the EU.

[1] Subsection 11 (1) in the Act in force until 1 January 2019. However, already in Judgment 3-3-1-3-12 of 2012, the Supreme Court found that subsection 11 (1) of the Personal Data Protection Act did not provide an independent legal right for processing data. Therefore, it is not a fundamental amendment but rather one for increased clarity.

1.3. What are personal data?

Personal data are any information relating to an identified or identifiable natural person. A person identifiable if they can be associated with the data. If the data cannot be associated or they are anonymised, i.e. processed in such a manner that identification is no longer possible, the data are not personal data and are not covered by data protection. Whether a researcher has to comply with data protection requirements or not, therefore, depends on the identifiability.

A person is **directly identifiable** if the data contain, for example, the name, personal identification number or another unique data unit. Direct identification is based on the available data and requires no additional data or knowledge. **Indirect identification** means that the link between a person and their data is not directly manifest but has to be created or inferred, for example, by combining several identifiers. Indirect identification may also be possible by combining different datasets. Such identification relies on assessment.

When a researcher starts to assess the identifiability of the data, it is not enough to look at the processed data. It is necessary to think a few steps further: what else can be done with the data if there is an interest in identifying or finding the people behind it? If there is a lot of background information, it can help make data attributable to the person, even if it appears anonymous at first glance.

The assessment of identifiability must take into account all reasonable and easily performed steps that can be taken to identify an individual. A person is not considered identifiable if the identification takes unreasonably much time, effort or means. Whether it is reasonable is assessed based on the resources required compared to the likelihood of identification. It is important to note that the line beyond which a person is no longer identifiable is not unambiguously clear and will be reassessed from time to time in light of new technologies and identification methods.

Generally, it is possible to distinguish three types of personal data related to research.

1. **Survey respondent data** are data collected from or about individuals, which need processing to achieve the objectives of the research; for example, recordings and transcripts of interviews, responses to surveys, observation and location data, results of experiments, health data, measurement results or other data relating to an individual.
2. **Contact details and other organisational information** are data relating to the participation of individuals in a survey; for example, the lists of respondents, their email addresses, telephone numbers, data concerning the location and time of the experiment, interview or another type of data collection, and written consent forms. These data are collected in the course of the research but are not strictly used to achieve the scientific objectives of the study. Nevertheless, they are personal data, and their processing must comply with data protection principles.
3. **Data on researchers**: several data on researchers may also be gathered during the study, such as the general data and contact details of researchers, data on their workload, salary and work-related travel, or information on who collected and analysed the data, when and how. Biographies submitted in research project proposals also contain personal data. The data on many researchers are publicly available in ETIS or on the website of the research institution.

The university's data protection wiki includes a checklist covering all stages of data processing in research. Before starting a survey, it is useful to estimate whether all the conditions for the secure and relevant processing of personal data have been met.

A natural person whose data are processed is called a **data subject**.

1.3.1. Special categories of personal data

Besides personal data, Article 9 of the GDPR distinguishes **special categories of personal data**:

- data revealing racial or ethnic origin, political views, religious or philosophical beliefs or trade-union membership,
- genetic data,
- biometric data that are used to identify a natural person uniquely,
- health data, or
- data concerning a natural person's sex life or sexual orientation.

Generally, the processing of special categories of personal data is prohibited, i.e. the **processing limitation** applies. Such data may only be processed based on a person's consent or other exceptions provided in Article 9 (2) of the GDPR. There must also be a legal basis for processing.

Examples of the use of special categories of personal data in research

Special educational needs. When researchers want to study students' special educational needs, it may be difficult for them to determine whether these are special categories of personal data. That is certainly the case when special educational needs arise from health reasons (for example, linked to a medical diagnosis or disability). When, however, the special needs are a talent or a communication or learning disability, they may not be related to the child's health. In this case, it is not a special category of personal data. Nevertheless, special educational needs can be considered a more sensitive data category than usual, especially for children.

The final assessment depends on the specific data processing and its purpose. If the aim is just to make generalisations and the researcher is not interested the cause of the special need, it is not a special category of personal data. For example, if students A and B in one class allegedly have (an unspecified) special educational need, but students C, D, and E do not, the fact of the special need is taken as such, and it is not considered a special category. However, if the research focuses on the performance of pupils in connection with a specific and explicitly formulated special need (for example, how speech impairment affects learning motivation or what kind of support pupils with a physical disability need), it is processing special categories of personal data.

If researchers cannot fully control the amount of data collected (for example, the researcher does not know what the interviewee answers or what is written in response to an open-ended question), they may end up collecting special categories of personal data even if the original purpose was to investigate special needs (such as talent) that do not fall under the special categories. It is therefore advisable, as a precautionary measure, to treat all special educational needs as special categories of personal data, especially if it cannot be excluded that the research will also analyse the causes of the special needs. If it is known that this will not be done, it should be clearly stated in the survey plan and the information given to the respondents. Researchers can also formulate data collection questions in a way that does not encourage anyone to share health data.

Health indices. Indices and other complex measures that can be used to draw conclusions about a person's health are considered health data. They should therefore be regarded as special categories of personal data, and the same applies to calculating the indices. For example, if the study aims to calculate a body mass index, associate it with a person and thereby obtain new information about the person's health, the calculation of the index must be considered processing of a special category of personal data. Weight and height do not belong to special categories of personal data.

1.3.2. IP addresses

In its judgment, C-582/14 of 19 October 2016, the European Court of Justice ruled that dynamic IP addresses that change with each connection to the internet constitute personal data. This interpretation is based on the GDPR's definition of personal data as "any information relating to an identified or identifiable natural person; an identifiable natural person is one who can be identified, directly or indirectly".

The European Court of Justice notes that "a dynamic IP address does not constitute information relating to an 'identified natural person', since such an address does not directly reveal the identity of the natural person who owns the computer from which a website was accessed, or that of another person who might use that computer" (p. 38). However, an IP address makes it possible to **identify a person indirectly**. The European Court of Justice explains that all the information enabling the identification of the data subject does not have to be in the hands of one person (p. 43). That means that the internet service provider may be asked to provide additional data, after which it will be possible to identify the person.

In addition, it should be assessed how **reasonable and likely** the possibility of indirectly identifying an individual is when combining the different data. According to the court, it is, for example, reasonable and likely that an internet service provider will transfer its customer's data to a competent authority (for example, in the case of cyberattacks) to protect the rights of individuals or comply with legal obligations. It is also possible that a person who knows the dynamic IP address of a potential infringer, in order to protect their rights or to comply with legal obligations, to take legal action before a court or other competent authority that can request the necessary data from the internet service provider to identify the infringer behind the IP address.

Reasonable and *likely*, therefore, include the possibility that a person is identified indirectly by several institutions or persons working together. An indirect identification is not reasonable and unlikely if it is prohibited by law or practically impossible and requires a disproportionate effort or cost (p. 46). Consequently, the researcher should be aware that if the respondent's IP address is stored together with the data collected, for example, through an online survey or another online service, the respondent could be identifiable from the perspective of the data protection law. Various survey platforms allow researchers to configure the survey in such a way that the respondent's IP address or other technical information, which would facilitate the respondent's identification, is not collected (see 3.4.5).

1.4. What is personal data processing?

Processing is a general term in data processing that means any operation performed on data. It includes data collection, storage, copying, modification, systematisation, retrieval, transmission and destruction. The processing of personal data goes on from the moment the data is received until it is destroyed and includes all operations in between.

Processing also includes the anonymisation of personalised or pseudonymised data, which, together with the prior collection of data, must comply with the general data protection principles.

The GDPR distinguishes between three responsible roles in the processing of personal data.

1. A **controller** is an institution that determines the purposes and means of processing, i.e. controls the processing of personal data. A controller is usually a legal person in which personal data processing occurs. As an organisation, the University of Tartu is the controller; the researcher is responsible for everything they do with personal data in their work. The researcher sets the data processing objectives and says what data they collect and by what means.

   The fact that the university may be the personal data controller does not mean that every university employee can access personal data. Such access must be needs-based and limited to researchers involved in the specific project or stage of research. In the case of sensitive data, such as children's data or special categories of personal data, access should be further restricted to researchers who absolutely need to process the personalised data.

2. Where decisions on personal data are made jointly by several institutions, such as the university and other research institutions, they are **joint controllers**. Joint responsibility means that requires that the cooperating institutions determine the purposes and means of processing together. For example, in EU-funded projects, the project partners may each be responsible for their own activities or act as joint controllers, depending on the division of work and decisions.

3. A **processor** is an individual or an institution that the controller has authorised to process personal data based on a contract. A processor works on behalf of the controller. As they cannot determine or change the purposes and means of processing personal data, processors are not controllers. A person working at the university under an employment contract is not a processor because the university cannot authorise itself to do the processing. However, a university researcher, who has been hired by another public or private body under a separate agreement to carry out analysis or expert assessment, can be a processor.

Usually, researchers are not processors or controllers: through their duties, they fulfil the obligations of the university as a controller. For example, if a researcher is the principal or responsible investigator in an international project, the university is the controller or joint controller – depending on the agreement between the research institutions – throughout the project. It does not mean that the researcher has no responsibility. It is a good practice to appoint a **researcher** who is **responsible** for the data processed in the specific research study and must ensure the accuracy of data processing. Their task is to ensure the confidential and secure data processing and provide guidance to researchers processing personal data.

Both the responsible researcher and the controller must follow data protection principles, but they report to different authorities. The controller, i.e. the university, reports a data breach to the Data Protection Inspectorate. The responsible researcher reports to the employer, i.e. the university, and must also notify the university's senior specialist for data protection. The university must assess the potential risks associated with the processing of personal data and take measures to mitigate them, while the researcher must carry out the risk analysis.

Thus, it depends on the agreement between the researchers and the university as to which staff member is responsible for data protection issues. The university's documentary procedure rules provide that the responsibility for personal data processing lies with the specialist for data protection, the head of the structural unit, and the employee processing personal data. Each employee who processes personal data must ensure data integrity and confidentiality. The head of the structural unit must ensure the registration of all personal data processing activities.

A **third party** means an individual or institution other than the data subject, controller, processor and persons who work under the direct authority of the controller or processor. To put it more simply, it is a person who does not have a clear role in processing personal data. In the case of a research paper, third parties may include, for example, the researcher's family members, an opponent or reviewer of the article, and employees of the publishing house publishing the article.

In addition, there are **recipients** – individuals or institutions to which personal data are disclosed. This role is situation-specific. The recipient may be a controller, processor or a third person.

Each institution, including the university, establishes its **data protection policy** which every person involved in research, from a student transcribing an interview to a professor leading a research project, must follow. The conditions set in the policy must be observed regardless of the purpose of processing. At the university, the same policy applies to personal data related to human resources, academic affairs and research. The purpose of research does not diminish the need to protect personal data or respect individuals' rights.

1.5. What are data protection principles?

All data protection legislation is based on the principles of the processing of personal data set out in Article 5 of the GDPR ("data protection principles"). These were in force before the GDPR was adopted, but their interpretation and implementation have slightly changed.

1.5.1. Personal data processing is lawful and fair

The processing of personal data is lawful when it has a legal basis. In its absence, the processing of personal data is unlawful.

Fair means that the interests and rights of the data subject must be considered and not unduly prejudiced. Even if processing is lawful, it may disproportionately harm a person and therefore be unfair.

1.5.2. Personal data processing is transparent

Transparency means that data subjects understand what is being done with their data and how, and that the respective information is clear and easy to find. For example, the obligation to provide information and to publish data protection conditions results from the transparency principle.

1.5.3. Personal data is processed for the intended purposes

The purpose limitation principle means that before the data are collected, the purpose of processing the data must be legitimate and explicitly stated. For example, it is not correct to refer to "research paper" or "research study" as the purpose; the specific end result of a project or study should be pointed out. However, Recital 33 of the GDPR on consent admits that it is not always possible to fully identify the purpose of personal data processing for scientific research purposes at the time of data collection. Therefore, it is permissible to state the field of research or part of the research project.

As a general rule, data may not be processed for purposes other than those specified in advance, but the GDPR provides for an exception to this principle in scientific research: data collected initially for other

purposes may be used in later research. Subsection 6 (1) of the Personal Data Protection Act lays down the obligation to pseudonymise data before their transmission for research.

It should also be noted that the purpose of academic work only covers activities directly related to the research, but the work also involves other activities in addition to scientific research. For example, if personal data are used for publications, teaching, scientific conferences, business, in the application of research results or science communication, they require their own purpose and legal basis.

It is difficult to assess the purpose limitation without knowing what data are being collected. Therefore, it is good practice to define the purpose by specific data subjects or types of data. For example, the university's data protection policy describes the purposes and legal bases for processing by types of data.

1.5.4. Personal data processing is minimal

According to the data minimisation principle, as few as possible data – only data necessary to achieve the purpose of the research – should be collected and processed. Collecting excessive data may be unlawful if there is no evident need based on the purpose. Therefore, to comply with the minimisation principle, the researcher should carefully consider what minimum data are needed for the purpose.

Example

The date of birth gives more information about an individual than the year of birth; the year of birth tells more than the age in years, and the exact age says more than an age range. If, for research analysis, the researcher wants to divide respondents into cohorts according to age ranges, it is against the minimisation principle to ask for a person's date of birth or exact age. The minimal solution would be to ask the respondents to which cohort they belong.

1.5.5. Personal data processing is based on accurate data

The principle of fairness implies that only correct data may be processed. They must therefore be checked, rectified, and, where necessary, updated or deleted. Data subjects have the right to demand the rectification of their data that are incorrect (see 2.9.3). This right can always be exercised, and no exception is made for research.

However, inaccurate data relating to an individual can also be considered personal data. Therefore, the data protection principles must be applied equally to all data, regardless of their veracity.

Example

If the contact details of the respondents to a longitudinal survey change, they should be corrected based on public database query results. While a request for contact details is justified in the light of the need for the research, questions may arise as to the legal basis for such a request – do researchers have the right to obtain contact details from the database, and does the controller of the database have the right to provide them to the researcher? Thus, the requirement to ensure data quality raises some controversy.

For clarity, it would be helpful to ask survey participants to give explicit consent to be contacted later or update their contact details based on public databases.

1.5.6. Limitation on the storage of personal data

Generally, data can only be stored until the purpose is fulfilled. After that, they must be deleted or anonymised. After the original purpose has been fulfilled, data may be retained if the processing is carried out in the public interest for archiving, historical or scientific research or statistical purposes (see also 4.1).

1.5.7. Personal data processing is secure

Security means that the availability, integrity and confidentiality of data, i.e. protection against unauthorised processing, must be ensured. Ensuring security is necessary to take various technical and organisational measures. Technical measures include, for example, data encryption; organisational measures include, for example, granting access rights to researchers or storing data on a single server rather than on each researcher's personal work computer.

1.5.8. Data protection by design

Data protection by design requires the controller and processor to integrate all data protection principles and work processes. This means that researchers must pay ongoing and continuous attention to data protection issues at all stages of data processing, starting from the research planning.

Data protection by design is linked to the concept of privacy by design, which has evolved from the principles of developing information and communication technologies and emphasises the importance of thorough protection of privacy and personal data. That, in turn, has led to the values by design or ethics by design approach, which emphasises the importance of considering human values when designing activities and work processes.

Example

When using the principle of data protection by design, the person is not asked for one single informed consent but for separate permissions for processing personal data used in the survey, reporting incidental findings, aggregating different datasets, participating in a follow-up survey, using personal data in further research, and for the retention of personal data after the planned end of the study. However, if consent is asked for more than one purpose at a time, the person must have the opportunity to opt out of some of the purposes.

1.5.9. Data protection by default

Data protection by default requires researchers to give preference to solutions that offer a higher level of protection of the individual's privacy if there is a choice. For example, when publishing interview-based survey results, if it is possible to choose whether to disclose the names of interviewees or keep them confidential, the default preference should be keeping them confidential. If it would allow researchers to achieve the same objectives in the data analysis phase, they should use pseudonymised data.

The principle of data protection by default is crucial when asking individuals for their consent to the processing of personal data, where the default answer is "no", and the participant must actively confirm their consent (see also 2.3).

1.5.10. Pseudonymisation and anonymisation

According to the GDPR, **pseudonymisation**[2] is the processing of personal data in such a way that identifiable information is removed and replaced by a pseudonym, such as a code or another identifier. However, such data are still personal data because they can be converted back to personalised data. Pseudonymisation is an additional safeguard that protects the data subject's rights but does not release the researcher from the responsibility to comply with data protection principles (see also 3.3).

Anonymisation is the processing of personal data in a manner that it is no longer possible to identify an individual, directly or indirectly, by any reasonable or likely means. While the data processing methods of anonymisation and pseudonymisation may be similar, the main difference is the *irreversibility* of the processing: pseudonymisation is reversible, but anonymisation is not (see also 3.3).

Data about people that cannot be linked to specific individuals is called **anonymous data**. The use of anonymous data is not regulated by data protection. The researcher has to assess whether it is still possible in any way to link anonymous data to a person or whether or how likely it is in the future.

1.6. What is a legal basis?

The processing of personal data is lawful only if there is a legal basis mentioned in Article 6 of the GDPR. These bases with brief explanations are listed in the general guidelines for data processors (pp 7–8) compiled by the Data Protection Inspectorate.

Depending on the study, the researcher has to choose which legal basis to use.

Consent is the most common legal basis in research. It supports the autonomy of the people involved in the study and ensures that their participation is voluntary. The consent must be informed, so there are various additional requirements for asking for consent (see 2.3).

Task carried out in the public interest is a legal basis for public research institutions, to which this function is assigned by law. For example, the Archives Act has tasked the National Archives with carrying out archival research and publishing it. Consequently, the task in the public interest is a suitable legal basis for the National Archives to conduct research involving personal data.

A task in the public interest is not an appropriate legal basis for private research institutions whose research activities are not subject to specific legislation. However, a private research institution may be commissioned or otherwise authorised by a public authority to do research, the legal basis of which is a task in the public interest.

Legitimate interest is a flexible basis that allows for a needs-based assessment of the importance of different opposing interests. The controller must weigh the legitimate interest against the interests of potential data subjects. Therefore, it is not sufficient that the research institution has a legitimate interest in the research or that the study is in the public interest – the interests must be overriding, and the potential harm to the interests and rights of data subjects must be minimised.

[2] GDPR uses the term *pseudonymization*.

Public authorities can use legitimate interest as a legal basis in very limited circumstances (e.g., employment certificates, photos on the intranet, use of cameras). However, private research institutions or research and development companies can do so.

In addition, the GDPR provides for three other legal bases – compliance with a legal obligation, performing a contract with a data subject and protecting an individual's vital interests – which are not relevant in the research context. In exceptional cases, however, these bases may prove necessary. For example, according to section 27 of the Child Protection Act, all persons who know of a child in need of assistance are required to report it. Therefore, if the researcher has contact with children or families in the course of the study and has reasonable grounds to suspect that a child needs help (for example, a victim of domestic violence), the researcher must notify the local government. In this case, the legal obligation is the legal basis for disclosing the child's personal data. Even if, rarely, such a need may arise while carrying out research, it is not the processing of personal data for research.

1.7. How is the Estonian Code of Conduct for Research Integrity connected with data protection?

In the Preliminary opinion (p. 12), European Data Protection Supervisor has expressed the view that exceptions to data protection apply only to research carried out ethically. Therefore, when processing personal data, it is necessary to respect the principles of both research ethics and data protection.

The European Code of Conduct for Research Integrity was adopted in Europe in 2017. It is a "guide to researchers in their work as well as in their engagement with the practical, ethical and intellectual challenges inherent in research". The code describes the fundamental principles and violations of research integrity.

In Estonia, the general principles of research integrity have been agreed upon in the Estonian Code of Conduct for Research Integrity, which deals with the same issues as this guide. The ethical and legal obligations largely overlap, but in some individual matters, more ethical obligations may be imposed on the researcher, while in other cases, more exceptions may apply. The following chapters describe the Code of Conduct for Research Integrity, which is directly related to the processing of personal data, particularly consent, information, access and data retention.

2. Planning research

This chapter explains the main issues related to personal data processing to be considered in the stage of planning your research. It deals with the important issues of the lawfulness of personal data processing, and research carried out based on consent or without consent. In addition, it explains the requirements for processing special categories of personal data and the rights of data subjects with regard to their data.

Data protection needs to be addressed throughout the research process and after its completion. On the one hand, in line with the principle of **data protection by design**, it should be considered at the project proposal stage when the tasks and division of work are not yet in place. On the other hand, data protection issues may arise years after the end of the project, for example, when the data collected for the original purpose are to be used in a further study. If it is clear already that the dataset is valuable or it could serve as a basis for another research project, consent should be asked from data subjects before the data are collected for the original work.

As data protection issues overlap with those of **data management**, it is recommended to deal with both simultaneously. For example, if a project involves pseudonymisation of personal data for security, it is possible to plan it early on: how to do it, who will have access to the key or the data, and what will be done with the key at the end of the project. All these decisions should be written down in the data management plan (see 2.2.1).

2.1. Where to start when personal data are to be processed in research?

First, an overview of the planned research and the personal data to be processed must be prepared. One should identify the requirements, constraints, and risks to be addressed in relation to the personal data collected for research. The earlier problems with personal data are identified, the more time there is to deal with them.

Below is a checklist with questions to consider when planning your research. Before answering, it is advisable to consult the funder's terms and conditions, as these can often coincide with the questions here, although in different wording, order or detail. The European Commission's Ethics and Data Protection Decision Tree can be helpful.

It is recommended that answers are documented immediately. If a data management plan is used for the research study or project (see 2.2.1), it makes sense to write the answers to the questions in the plan. In addition, the written answers may be helpful when drafting the terms and conditions of data protection, a consent form, an approval request to the ethics committee, a funding application or other documents related to personal data.

CHECKLIST ON USING PERSONAL DATA IN RESEARCH

1. What is the purpose of the research study?

The purpose must be formulated as precisely and specifically as possible. In the most general terms, the purpose should provide people with information about what is being researched or what the study intends to achieve (see 2.12).

2. Do I need to process personal data?

If it is possible to achieve the purpose of the research study without processing personal data, that option should be preferred. It is useful for several reasons: it reduces the resource costs associated with the collection and proper management of personal data, spares those involved in the survey, reduces the risks arising from the misuse or insecure management of personal data and, in turn, avoids harm to people's privacy arising from these risks. If the conditions are equal, a solution that requires less processing of personal data should be chosen – this is in line with the minimisation principle (see 1.5.4).

3. What personal data need processing?

One should only collect data necessary to achieve the purpose, i.e. no data should be collected just in case. Personal data required for the research should be documented together with the purpose for which they are used – this helps create a better overview and thereby better assess whether all data are strictly necessary.

The personal data used in a study should be presented by type or category. This information can be used in the data management plan and the overview of processing personal data in research projects.

4. Do I process special categories of personal data?

In the planning stage, it must be clarified whether the data to be collected include special categories of personal data (see 1.3.1). If you plan to process special categories of data, it is necessary to request approval for that from the appropriate ethics committee. (see 2.2.4, 2.8.5 and 2.11).

5. Who are the data subjects of a research study?

An overview of the data subjects should be presented by type or category. In the absence of a general classification of people, the same terms as used for sampling, for example, "50–60-year-olds", "basic school pupils", or "minors", should be preferred. If there is no need to distinguish between data subjects based on specific characteristics, you may use generic terms such as "individuals involved in the study" or "respondents" (see 2.10, 2.15 and 2.16).

6. What is the method of data collection?

The main question is whether data are collected directly from individuals or whether data collected previously for other purposes, or secondary data, are used. For secondary data, it is necessary to determine where the data originate from, what agreements have been made with data holders and how the secure transfer of data to investigators is ensured (see 2.8).

Irrespective of the data collection method, people must be informed of processing their data (see 2.8.2).

7. What is the legal basis for data processing?

One of the legal bases provided in the GDPR (see 1.6) must be present to ensure that the processing of personal data is lawful. Different stages of a research project may have different legal bases, but you cannot choose more than one in the same phase. In most cases, the question is whether the processing of personal data is carried out with the person's consent (see 2.3) or on another legal basis (see 2.4).

8. Where and how are the data stored?

It is necessary to think about where and how the data is stored, who has access to them and under what conditions. For example, it is necessary to decide whether the data are stored on a university server, staff members' computers, a cloud service, etc., and whether the environment is located in the European Union or a third country. If archiving of data is planned, one should decide in good time how and where the data will be stored (see 2.2 and 3.1).

9. How long are personal data stored?

It must be agreed beforehand in which form and for how long the personal data used in the research will be stored. There is an exception for scientific data, allowing to keep them beyond the original deadline, but not indefinitely. It is common practice to store data for 5–10 years after the end of a project or research study so that it would be possible to verify the results (see 4.1).

Backing up and long-term storage of data is closely connected with the management of research data. For further information, read the materials published by the University of Tartu Library, Research data management and publishing and Data management plan.

10. What security measures are planned?

Security measures can be technical or organisational. Technical security measures concern data processing equipment and environments. Organisational security measures primarily include work procedures, physical access restrictions (access cards), locking of premises and equipment, data management plans or registration of processing of personal data. Also, encryption, pseudonymisation and anonymisation (see 3.1, 3.3 and 3.4) can be considered security measures.

11. Who are the recipients of personal data?

All recipients of personal data, i.e. the persons to whom personal data are transferred, should be indicated. They may be institutions or individuals involved in the project or external ones (see 4.3).

12. Are personal data transferred to third countries?

Transferring personal data outside the European Union involves the obligation to ensure that the country of destination provides adequate personal data protection. The data management plan and data protection policy should also describe the composition and format of the data to be transferred to third countries (see 3.2).

2.2. How much personal data needs to be documented in research?

Before starting the research, it is necessary to find out what kind of personal data processing documentation is needed for the planned study. The relevant possible and required documents are listed below. In addition, it may be necessary to describe the personal data processing elsewhere, for example, in applications and reports submitted to funders.

2.2.1. Data management plan

A data management plan is a tool for describing the data and the work done with the data. The planning starts with the most general answers about where and how the data will be obtained, what types of

data will be used and how they are related, what data formats will be used, how much data will be stored, what software will be used and where, how and for how long the data will be stored.

The data management plan helps describe data to make them findable, accessible, interoperable and reusable in the interests of open science (FAIR principle). The planning, however, also enables, in good time, to identify the potential problems, obligations and requirements associated with personal data processing. For example, it gives an overview of which part of the data is personal data, whether special categories of personal data are processed, whether data previously collected for other purposes are used, how data confidentiality and integrity are ensured, or with whom the data will be shared.

Although there is no general obligation at the university to draw up a data management plan for every study, systematic data management is becoming common practice. It may also be required by funders of the research, such providers of the Horizon 2020, the European Research Council and the Estonian Research Council grants. As the data management plan provides a systematic overview of all the data to be collected and analysed, it is reasonable to compile it at the same time as writing the review of personal data processing and, if necessary, an application to the ethics committee.

See also:

- University of Tartu Library's guidelines on creating a data management plan
- University of Tartu Library's course on research data management
- University of Tartu Library's examples of data management plans
- DMPonline tool

2.2.2. Data protection policy

All institutions that process personal data must publish their data protection conditions. It may also be necessary to draw up separate data protection conditions for more extensive research projects. For example, the university's data protection policy provides general information, but data processing in the context of a large-scale research project should be described separately. In international research, this is usually done.

Consent-based surveys present many of the data protection conditions on the informed consent form. As the transparency principle requires that data processing information should always be available and easily accessible to individuals, the same information should also be published on the research project's or the controller's website (see 2.7).

Read more:

- Data Protection Inspectorate's general guidelines for data processors, annex 3 "Data protection checklist"
- University of Tartu data protection policy

2.2.3. Overview of personal data processing

Each controller and processor have the obligation to maintain records of the processing of personal data, according to Article 30 of GDPR[3]. Generally, making a separate overview for each study or project is not justified or necessary. However, the university may not actually know how exactly personal data are processed in a large-scale study or project. Also, the university may share the responsibility for personal data with numerous other research institutions. Therefore, it may be necessary to write an overview of the processing of personal data for a single research project, especially if it involves the processing of sensitive data or the use of higher-risk processing methods.

Read more:

Data Protection Inspectorate's general guidelines for data processors, chapter 4 "Overview of personal data processing"

2.2.4. Ethics committee's approval

Several Estonian law acts have tasked ethics committees to assess whether the proposed research study complies with data protection requirements. The ethics committee's approval is either mandatory or voluntary, depending on the research. In an approval request, the researchers must describe, among other things, what personal data are processed, on what legal basis, and how and for how long (see also 2.13).

2.2.5. Data protection impact assessment

If the research data processing poses a high risk to people's rights and interests, a data protection impact assessment may be prepared to protect them. The data protection impact assessment is mandatory for the controller. The researcher should contact the data protection officer if they think an impact assessment might be necessary for the planned research.

In the case of collaborative projects, it should be explicitly agreed on which partners are responsible for carrying out the impact assessment and how other partners are involved. In addition, it is worth bearing in mind that impact assessments are carried out in different ways in different EU countries. In the case of international projects, it would therefore be wise to discuss and agree beforehand on how and by whom the impact assessment should be carried out.

A data protection impact assessment is a specific obligation, which does not mean that other types of risk assessment are not necessary. Depending on the situation, a data security or ethical risk assessment may also be necessary (see also 2.14).

2.2.6. Informed consent

Consent is one of the possible legal bases for processing personal data in research. An informed consent sheet must contain the most relevant information on the processing of personal data. Sometimes it is necessary to make several versions of the same information sheet; for example, one for adults and one for children. It may also be necessary to translate the information into different languages.

[3] Instead of the term *recording of processing activities* used in Article 30 of GDPR, the terms *personal data processing overview* and *overview of personal data processing* are preferred in Estonia. These have also been used in this guide.

The informed consent form with a person's consent is an official document that must be appropriately kept. A researcher may need to provide evidence of a data subject's consent if, for example, the person contests the processing of their data. Also, the ethics committee or funders may request access to the informed consent form to assess its compliance.

The consent can be withdrawn, and such withdrawal must also be documented (see the following subchapter).

2.3. What must the consent include?

The purpose of consent is to give data subjects as much control over their data as possible. It is, therefore, not a suitable legal basis for research studies where the individual's ability to control the processing of their data is limited.

According to the GDPR, consent to participation in a study is deemed valid only if it has been actively expressed and clearly confirmed (for example, by writing the word "yes", ticking a box, or signing). Therefore, the data subject must give **voluntary**, **informed**, **specific** and **unambiguous** consent to process their personal data (opt-in). The opposite situation, where consent for collecting personal data is implicitly presumed, and the participant has to do something to opt out, is prohibited and contrary to the GDPR.

A selection of sample consent forms can be found on the University of Tartu intranet.

Read more:

- The European Data Protection Board's guidelines 05/2020 on consent under Regulation 2016/679/EU
- Data Protection Inspectorate's general guidelines for data processors, annex 2 "Consent checklist"

2.3.1. Consent must be freely given

A person must not be influenced to give consent. Consent is not voluntary if the person has no real freedom of choice. Enticement with gifts, money or other benefits, persuasion or coercion is not allowed.

Consent is not deemed as freely given if

- there is an evident power or dependency relationship between the giver and the seeker of consent; for example, when a lecturer asks for consent from a student or an employer from an employee;
- giving consent is a precondition for providing a service or using a benefit or opportunity. For example, participation in professional conferences or seminars cannot be made conditional on consent to processing personal data (taking pictures or broadcasting), without which access to the event is denied. The same applies to open data repositories: for example, on a website giving access to open data, consent to cookies must not be required to access the data. In the case of a research study, however, it is rather difficult to imagine a situation where data collecting would be linked to an activity or service of interest to the data subject;
- the withdrawal of consent would have adverse consequences for the individual. It is, therefore, necessary to ensure that both the giving of consent and its withdrawal are voluntary.

2.3.2. Consent must be informed

Being informed means knowing and understanding what one agrees with. According to the GDPR, the information about consent must be given in clear and plain language. Complicated scientific or legal terms should be avoided.

On the consent form, the individual must be given information about the processing of their data. Otherwise, the consent given by the data subject cannot be considered informed consent and is invalid. The consent form must explain the entire data processing process from the beginning to the end (data collecting, analysis, transfer, and storage). This inevitably implies that *all* the information required by the GDPR must be presented to the individual when seeking consent:

- the name and contact details of the controller and the controller's representative;
- the contact details of the specialist for data protection;
- the purpose and the legal basis of processing personal data;
- the recipients of personal data, i.e. those to whom the data are transferred;
- the retention period of personal data;
- the right to withdraw consent and other rights related to data;
- information on the transfer of data to third countries;
- information on automated decisions and profiling of natural persons.

Being informed therefore requires a compromise between two conflicting interests. On the one hand, there must be enough information to give the reader an overview of personal data processing. On the other hand, the information must be simple and clear to ensure it is understood.

The information may be provided in any form. It may be written text or, for example, video, audio, animation, images or icons. Giving several versions of the information may be useful – one shorter and simple, the other more extensive, detailed and text-based.

2.3.3. Consent must be specific and unambiguous

Specificity means that the purpose of processing personal data has been clearly stated. Formulating a precise purpose can be difficult in research as it is often unclear at the beginning what data will be processed, how they will be processed, and what are their future uses. Therefore, a certain concession has been made: the purpose of the research should be as specific as possible at the time. Although the purpose is less specific, it can be remedied by greater transparency throughout the research (for example, by keeping data subjects informed of the progress of the project) or repeatedly asking for consent (for example, by asking for new consent after certain stages of research).

The requirement that consent must be unambiguous is closely connected to the principle of transparency – it means that consent must be unambiguous and clearly formulated, without any misleading or confusing statements. For example, if there are multiple purposes, the consent form must be drafted so that data subjects can choose which purposes they do or do not agree to (see also 2.12).

2.3.4. Consent to data processing must be clearly distinguished from other requirements and consents

The consent to processing personal data used in research is sometimes very similar to consent to participating in a survey. The giver of consent must understand that, for example, the consent to

participate in a clinical study does not automatically mean giving consent to processing personal data. Instead, the person must give double consent: one for processing personal data and the other for participating in the study. However, they are closely connected and if, for example, the person does not agree to data processing, they cannot participate in the study.

In addition, it is worth remembering that the consent to participate in a research study must also include information not required in the consent to data processing. The consent to participate must describe the purpose and organisation of research, the researchers, research institutions and funders involved, the expected societal benefits and potential harms of the study, and a description of how these are weighed against each other, the possible uses of the results, the risks to people and the measures envisaged to mitigate them.

2.3.5. It must be possible to prove consent

There are no specific requirements for the consent form other than that it must be verifiable, i.e. documented. Therefore, it should be in writing. Verbal consent is also suitable as long as the researcher has recorded it and it can be reproduced.

Consent does not necessarily need to be signed; therefore, an email is also acceptable. However, it must be possible to prove that the data subject has given consent. If problems arise with the processing of personal data and the research institution cannot verify that the consent exists, the Data Protection Inspectorate may consider the processing unlawful.

Documented consents are also personal data. In addition, they are official working documents of the university, and their retention time, manner and place have been agreed upon in the documentary procedure rules. Generally, they must be retained until the end of the data processing.

2.3.6. Processing must be limited to what is described in the consent

Consent is valid only under the conditions described in it. If the consent form does not inform an individual of certain processing activity, there is no legal basis for such processing. If the nature or extent of processing significantly changes during the study, new consent must be obtained to continue processing the personal data.

In the case of a follow-up study, it is necessary to obtain new consent at the time of the new data collection, even if the person was informed of the follow-up study at the time of the initial consent request.

2.3.7. Consent must be easy to withdraw

A person must be able to withdraw consent, and this should not be unduly complicated. Otherwise, the consent is void, and the personal data processing is unlawful.

If a participant in a study wants to withdraw their consent, they should file the respective request in the form of a digitally signed statement or using other methods of identification. The data processor must ensure that the person withdrawing consent is the same person who gave it.

The withdrawal-related communication should not be limited to withdrawing the consent; it is also good practice to explain to the data subject what will happen to their data after the consent is withdrawn. For example, it should be reiterated that the processing of data during the validity period of the consent was

lawful, and cannot be withdrawn. The data processing ends of the moment the person gives notice of withdrawing the consent.

2.4. What to consider when personal data are processed without consent?

According to section 6 of the Personal Data Protection Act, personal data can, in certain cases, be processed without a person's consent, relying on another legal basis, for example, performing a task in the public interest. There is, however, little clarity about situations where the legal basis for the research is a public interest task, and it is therefore not possible at this stage to make specific recommendations for research without consent. Below, a few general recommendations are given, which should be taken into account when carrying out research in the public interest without consent.

2.4.1. Data must be pseudonymised or additional requirements met

According to subsection 6 (1) of the Personal Data Protection Act, personal data may be processed for research purposes without the person's consent in a pseudonymised format or a format providing an equivalent level of protection. In particular, the pseudonymisation requirement concerns secondary research, for example, where the researcher wishes to use information from a public database or personal data previously collected for another purpose and the legal basis is a public interest task.

As an exception, personalised data may be processed without consent if all the conditions specified in subsection 6 (3) have been met.

- **The purpose of research can only be achieved with personalised data**

Personalised data may be used if the purpose of research is drawing conclusions about particular individuals, the information of interest is very closely linked to the specific individuals, or the research relates directly to the interests and rights of the specific individuals. The researcher must always justify the need for personalised data and explain why they cannot carry out the planned research with the consent of the persons.

- **There is an overriding public interest in the research**

Public interest may be involved in both applied and basic research. Performing a task in the public interest implies that the planned research has a scientific or social value that the researcher must explain. The public interest must be overriding – for example, if the study involves a significant interference with people's rights and freedoms, it can be justified by a very high public interest.

- **The research must not damage the data subjects' rights**

The researcher must assess the potential harm to the data subject's rights. If there is no risk of harm, the planned research is very likely to be allowed. However, if there is still a risk of harm to the rights, it is necessary to take mitigating measures or find another way to carry out the research.

In most cases, it can be presumed that if all the criteria of section 6 of the Personal Data Protection Act are met and the reasons given are convincing, a public interest task can be established as the legal basis.

Preparing a justification and assessment under section 6 (3) is the responsibility of the researcher or the research institution. If the personalised data fall under a special category of personal data, the ethics committee's approval must be obtained.

2.4.2. Reference should be made to the legal provision

A task carried out in the public interest must result from the law. Therefore, for the sake of transparency, the public interest of the planned research should always be indicated, and reference should also be made to the legislative act specifying this public task.

The research mission of some public research organisations has been laid down in a separate law act under which the organisation operates. In this case, it is possible to refer to a specific provision. In other cases, the Personal Data Protection Act may be the only legal instrument which does not, however, guarantee that there is automatically a public interest in every planned research project. In this case, researchers need to provide further justification or explanation of the public interest.

2.5. How to ensure the lawful processing of personal data?

The processing of personal data in research is lawful if it has a legal basis and complies with the requirements of the GDPR, the Estonian Personal Data Protection Act and the data protection principles arising from them. Further recommendations can be found in the European Data Protection Board's guidelines 4/2019 on Article 25 of the GDPR, "Data protection by design and by default". The guidance below explains how to comply with the principle of lawfulness in research.

2.5.1. The legal basis is determined before the processing of personal data starts

The main question in research studies is whether personal data are processed with or without the person's consent. Processing personal data without consent always requires justification and an appropriate legal basis, such as a contract, legal obligation, or performance of a task carried out in the public interest (see also 1.6).

Example

A researcher studies the public material people have posted about themselves on social media. The researcher does not plan to seek people's consent knowing that it is not common practice when analysing material published on the media. Later, when the data have been collected, the researcher discovers that the people are identifiable from their comments, which can be found by text search. Even if the researcher does not publish the collected material, it has still been a case of collecting and analysing personal data without consent, which would require a legal basis. While the researcher may find that it is a public interest task, this should have been ascertained before starting the data collection.

2.5.2. Most appropriate legal basis must be determined

As the GDPR allows for a certain degree of flexibility regarding the legal basis, the possible and appropriate legal basis for the specific study may sometimes be unclear.

The legal basis must be clearly formulated, and it must not be misleading. People must not be given the impression that they have been asked for their consent when actually their consent is not the legal basis for processing personal data. If an individual withdraws their consent, the research institution must not

process their data for the same purpose on other grounds claiming, for example, that the research is a task carried out in the public interest and the data may therefore be processed without consent.

Example

A researcher requests data for his research from a public database. The public institution holding the data gives the researcher a pseudonymised dataset where identifiers, i.e. the information allowing direct or indirect identification of an individual, are encrypted. The researcher understands that these are pseudonymised data, and their processing requires a legal basis. Since in this specific public database, every Estonian citizen can determine their preference for the release of data for other purposes, including research purposes, the researcher considers consent to be the legal basis because the release of data is based on the person's free will. This assessment, however, is incorrect. The choice indicated in the dataset is necessary for the public institution to be entitled to release data for research purposes. However, such consent does not meet the conditions of the GDPR – for example, it is not clear how a person can withdraw the consent. Moreover, the researcher has not given people the necessary information to obtain their consent.

2.5.3. Related activities that may need a separate legal basis should be distinguished

Each processing activity should be clearly distinguished based on the purpose and legal basis. While a small-scale study only has one purpose (the anticipated result of the research) and one legal basis (consent), it is easy to comply with this recommendation. A large, long-term research project, however, may have more purposes and in that case, each purpose must be linked with a legal basis.

If, in the same research, the intention is to collect data directly from people and combine them with previously collected data that may be located elsewhere, a separate legal basis should be assigned to each operation even if they serve the same purpose.

Example

A research project uses personal data that cannot be anonymised without compromising their quality. Researchers still want to store them in a scientific data repository and make them available to other researchers. While open science is important, sharing personal data with other researchers is not necessary to achieve the goals of a particular research study. That is an independent purpose. One possible solution is to seek extra consent from the survey participants. In this case, they choose whether to allow using their data only for the specific study or future research.

2.5.4. People are given as much freedom of choice as possible

When planning research, researchers should give data subjects the opportunity to choose how they want to participate in the study and how their data will be used. It depends on the specific research how much autonomy and freedom to make their own decisions can be given to people. For example, in the case of an interview with an expert, it is possible to determine how or what they say is presented in the published text: whether the expert's name or a pseudonym is used, or a reference is made to the institution where they work.

Where it is possible to choose between several legal bases, consent should be preferred. In cases where asking for consent is impossible or would hinder the fulfilment of the research objectives, other legal

grounds may be considered. However, even then, the controller should ensure as much autonomy as possible for the individual – for example, a choice to prohibit using their data for other purposes.[4]

2.6. How to ensure fair processing of personal data?

Below are the recommendations listed in the European Data Protection Board's Guidelines 4/2019 on Article 25 of the GDPR, "Data protection by design and by default", and a brief explanation of how to comply with the fairness principle for processing personal data in research.

2.6.1. Processing of personal data should correspond with people's expectations

The description of the processing of personal data used in research must be as accurate as possible and include all the circumstances that could be reasonably expected to be of importance to people or likely to cause distress. Information about the research and personal data processing must not be misleading or create unjustified expectations in the data subject.

Example

It is explained to participants in the survey that their data and the interview contents are confidential, and no one other than the organiser can see them. After the end of the survey, a participant discovers in the published final results that an extract from the interview made with them has been published: although the participant's name has not been revealed, the words are recognisable. The person is disappointed and informs the author because they had understood that the interview would remain completely confidential. The researcher, however, explains that it is a common academic practice to quote extracts from interviews in qualitative analysis.

Such a situation may seem unfair to the participants as they do not need to know academic practice. The researcher's actions are not legal from the perspective of data protection, but the case illustrates how the individual's expectations and the researcher's assumptions may differ. Additional explanations about the purposes of the research and data processing would help prevent such situations.

In qualitative social science research with small, homogeneous samples, it is a common practice to show the research results to the respondents before publication. This helps get feedback and better understand the respondents' expectations. It also improves the transparency of the research, and ensures that people's expectations are aligned with work done for the study.

2.6.2. It must be possible to communicate directly with the controller

The possibility to contact researchers at any time helps ensure transparent, fair and reliable processing of personal data. That may be more difficult when data collection has been delegated to a processor (for example, a survey company) and their staff, who may not know in detail the purposes of the survey, the data processing conditions, or any other relevant information that people may ask for. In that case, the research team responsible for the survey should include information on how to contact them and provide their contact details. In addition, the contact details of the controller's specialist for data protection may be supplied.

[4] For example, in the e-population register, individuals can restrict the disclosure of their address to private companies, even though the disclosure is not based on consent.

2.6.3. Discrimination in the processing of personal data must be avoided

The non-discrimination principle is universal and does not concern data protection only. If someone has been discriminated, all the data processing may become unfair and, therefore, unlawful.

In the case of scientific research, it is not always clear which processing can be considered discriminatory. Certain groups in society may be under- or overrepresented in studies, depending on how easy or convenient it is to collect data from them. Such discrimination may diminish the quality of the study, as conclusions drawn from a biased sample may not be valid, especially if they should represent a cross-section of society. It is common practice to impose additional conditions on survey participants to obtain a representative sample; certain participants are favoured based on those conditions. For example, in threshold surveys, it is common practice to interview the youngest male at home first, as this segment of the sample is the most difficult to reach. Sample-based criteria may be considered non-discriminatory, as people get no benefits for participation in the survey.

2.6.4. Exploitation of people's needs or vulnerabilities must be avoided

People must not be manipulated or exploited for research purposes. In research, it is essential to ensure voluntary participation because any form of manipulation limits the person's ability to make independent decisions. Particular care should be taken when subjects are not completely free or do not perceive their actions or behaviour as being researched.

The issue of vulnerability and exploitation is closely linked to research ethics. From the data protection perspective, everyone has equal rights to the data about themselves, regardless of their vulnerability. The only exception is children: their data must be processed with particular care to their rights and to ensuring that the children's rights are not harmed.

2.6.5. Asymmetric power balance must be avoided

Power relations must be taken into account when assessing whether participation is voluntary because, in the sphere of influence of someone in a higher position, the respondent may feel pressured to disclose personal data. Therefore, it may be unfair for a lecturer to include in a survey the students studying in a course they teach and whose exam results depend on the lecturer. Similarly, voluntariness can be problematic when a doctor wants to involve his patients, or a manager of an institution invites staff members to the survey.

2.6.6. Processing of personal data is ethical

Personal data processing that does not comply with the ethical principles of the field of research may be regarded as unlawful. The importance of research ethics is underlined by the requirement to obtain approval from the ethics committee for processing special categories of personal data. It should be taken into account that a breach of the principles of research ethics could, in the worst case, lead to data protection prosecution.

2.7. How to ensure transparent processing of personal data?

Transparency means that the data subject knows and understands how the personal data will be used in the research. To achieve transparency, the survey participant must have access to information both before the research and during data processing. On the other hand, the amount and quality of

information to be provided is an estimated value, and a research study's required level of transparency is not explicitly defined.

Below is a comment on the European Data Protection Board's Guidelines 4/2019 on Article 25 of the GDPR, "Data protection by design and by default", and a brief explanation of how to ensure as much transparency as possible in research.

2.7.1. Provided information is clear, understandable and relevant

When informing a data subject, the researcher must avoid complex words and sentences, professional terms, and ambiguity and avoid misleading the data subject. The information provided must not be a voluminous mass of text which is difficult to read. A good solution is to present the information in stages: a brief summary is made of the most important information, with references to additional information that give a more detailed overview of personal data processing.

The information must be provided according to the target group. For example, when information is presented to children and adults, it may be necessary to present it with different levels of comprehensibility. In certain situations, the information may require simplifying.

2.7.2. Time and channel of information are appropriate

Various ways should be used to provide information, taking into account the data subject's needs. The information must be easy to find.

- If personal data are collected in the course of an interview, the most appropriate time for informing the respondent is immediately before the interview.
- If data protection information is sent to the person with the survey invitation, the information should also be available at the place of the survey before the collection of personal data starts. In addition to the information sheet, the information on personal data processing should be available on the website of the project or the university.
- If a researcher collects data through a social media platform, people should be informed, in addition to other channels, also through the social media platform.
- The most important information could be in machine-readable form, but the GDPR also allows to inform people orally if they so wish.

2.7.3. Information on the algorithms used is provided

The GDPR specifically covers automated personal data processing, which results in a decision about an individual or their behaviour based solely on automated processing and produces substantial effects (legal effects or consequences of comparable importance) on the individual. For example, it is unacceptable to make recruitment and financial decisions based on automated profiling.

If the researcher plans to use automated processing, such as machine learning algorithms, to make decisions or inferences about a person or their behaviour, it must be explained to the person. Explicitness is the ethical principle supporting transparency when using AI; any decision made without human intervention must be understandable to the survey participant. The person must also be told what the expected outcome of the solution is and what it will be used for.

However, the GDPR does not limit the general use of automated data processing. Machine learning methods, which aim to find significant relationships based on large amounts of data and numerous

attributes, do not lead to legal consequences and are not limited in any way. The same applies to making statistical inferences from generalised data, which does not entail the obligation to provide information on each algorithm underlying the statistical calculation.

Therefore, the need to inform depends primarily on the impact of the processing on the individual and needs to be assessed case by case.

2.7.4. In the case of joint liability, a clear distinction must be made of what for and to what extent each person is liable

If several processors are responsible for the same processing operations, their tasks must be clearly distinguished. The joint responsibility of research institutions should always be agreed on in a separate agreement, which can specify to what extent they carry joint responsibility and to what extent separate responsibility.

Example

If the university is responsible for the collecting and primary processing of personal data in an EU project, but data from all the project countries are sent to the co-responsible partner for aggregated analysis, such division of responsibility must be explained to the respondent. This way, the participants know which research institution controls the use of their data at each stage of the project and who they need to contact to exercise their rights. If the data subject does not understand which of the numerous research institutions on the list can access and hold their data, the information is not transparent enough.

2.8. What to consider when using secondary personal data?

Based on the origin of data and the collection method, a distinction can be made between primary and secondary data. **Primary personal data** are collected directly from the individual. In contrast, **secondary personal data** have been first collected for other purposes and are not obtained from the individual but from databases, archives or elsewhere.

Processing primary data for research purposes is generally easier and more straightforward, as the legal relationship is only between the data collector and the data subject, the legal basis is consent, and the controller bears all responsibility for processing personal data. Secondary personal data, however, are often processed without consent, and not only the data subject's but also the controller's obligations and interests must be taken into account. Therefore, processing secondary personal data is generally more complex, as several exceptions apply.

2.8.1. Secondary use may be compatible with the original purpose

If the new use of existing personal data is compatible with the original purpose for which the data were collected, it will generally not require a separate legal basis. Therefore, when using secondary data, it is necessary to assess whether the original purpose of processing and the new purpose are compatible. Although in research, Article 5 (1) b) and Recital 50 of the GDPR always presume consistency with the original purpose, it does not mean that any processing of previously collected personal data is automatically allowed in research. The controller must also consider other data protection principles and, on that basis, ascertain whether it is permissible to use the secondary data.

Although there is no single principle for assessing the compatibility of the original and the new purposes, the compatibility is presumed to be higher if the same controller processes the data for both purposes. Compatibility is lower or non-existent where personal data are transferred to a new controller, a significant difference exists between the original and the new use, special categories of personal data are processed, or the processing presents a higher risk to the persons' rights and freedoms.

If the new purpose is not compatible with the original one, a new legal basis for processing is needed (see also 1.5.3 and 2.8.4).

2.8.2. Providing information to the data subject when collecting secondary data

In the case of primary personal data, all relevant information is provided directly to the data subject either before or at the time of collection. In the case of secondary research, when personal data are not collected directly from the data subject, some exceptions to the obligation to provide information apply (see Article 14 of the GDPR). For example, there is no obligation to provide information if the provision of information proves impossible or would require a disproportionate effort (data were collected a long time ago or there is a vast number of data subjects). However, to protect the data subjects' rights and interests, the information must be made public, for example, on the website of the research institution or project. In such cases, the controller is not obliged to contact the data subjects as they can be expected to find the necessary information on their own.

2.8.3. Secondary data holders

Secondary use of personal data is favoured, for example, under purpose limitation in Article 5 (1) b) of the GDPR and under subsections 6 (1) and (3) and the explanatory memorandum of the Personal Data Protection Act. The fact that it is necessary to consider the obligations of the data-holding institution towards data subjects makes it more complicated. Data protection legislation does not impose the obligation to provide secondary personal data to researchers, but considering the freedom to conduct research and the European Union's commitment to open science, the freedom of information and the duty of public authorities to provide information, public authorities – including holders of public databases and registers – generally provide researchers with the information they request, provided that they meet all the respective requirements. The explanatory memorandum to the Personal Data Protection Act also supports it. Commenting on the provisions of subsection 6 (3), it states, "Research with ordinary personal data does not require the approval of the Data Protection Inspectorate or the ethics committee. Persons doing research or a similar activity who meet the conditions must be given access to information, such as databases."

The situation with public databases is more straightforward. Many more problems arise in the private sector, where information may be protected by a commercial secret. Thus, even when the researcher has carefully assessed the need for personal data processing, public interest towards the processing and the proportionality of the infringement of the data subjects' rights, it is always possible that the controller will still refuse to release the secondary data.

2.8.4. There must be a suitable legal basis for secondary use

The controller who collected the primary data may transfer them to a researcher or a research institution, and the researcher or institution may accept them for secondary processing if both parties have a legal basis.

- **Public interest task as a legal basis for secondary use**

When using the public interest task as a legal basis, assessing the public interest served by the research is necessary. Section 6 of the Personal Data Protection Act allows the processing of personal data without consent in research, but imposes additional requirements: pseudonymisation (subsection 1) or the use of personalised data in exceptional cases (subsection 3, see also 2.4).

Researchers must also prove public interest. There are no agreed forms or standards for that. Still, since public interest is also demonstrated to the research funder or the ethics committee, the ethics committee's approval is usually sufficient to convince the data holder of the public interest. However, the data holder may demand that the researcher proves the legal basis of the public interest task otherwise.

- **Consent as a legal basis for secondary use**

Consent may not be an appropriate legal basis for secondary use of data, because it is difficult, if not impossible, for researchers to obtain it from data subjects with whom they have no contact. However, if reasonably possible, the data holder could seek consent by making a one-off request or using another solution, such as the consent service. Consent would assure both controllers that releasing data for a specific research project is lawful. It would also give the research subjects more control over their data.

Alternatively, it is possible to seek consent for the secondary use of personal data at the time of collecting the primary data already. In this case, the institution holding the data can be sure that it has the right to provide personal data to researchers or research institutions it trusts. At the same time, the promises given to individuals in the original consent must be respected. For example, if it was promised at the time of the initial data collection that the data would be kept for five years after the end of the project and then destroyed, the secondary use of the data must also fall within that period (see also 2.3).

2.8.5. Approval of the ethics committee is required for special categories of personal data

If special categories of personal data are required for secondary use, and the processing is not based on consent, the approval of the ethics committee is required pursuant to subsection 6 (4) of the Personal Data Protection Act. Release of data from the health information system or the biobank must be coordinated under other law acts. The ethics committee's approval is just an additional safeguard that does not provide a legal basis for secondary processing. A task in the public interest could be the legal basis for a research study without consent (see also 2.13).

2.8.6. Contract may be required for the transfer of data

It is possible to obtain personal data by entering into a contract with the data holder, in which the conditions, purposes and time limits for processing personal data are agreed. That allows the data holder to verify that the released data are correctly processed and lay down conditions for destroying or long-term storing the data, among other things.

The research institution that receives the personal data is the controller for further processing. Thus, both the research institution and the researcher, as a representative of the institution, are responsible for complying with all requirements of the GDPR, even in the absence of a separate contract. The researcher and research institution will not have the obligation of data protection only if the data are anonymised before the release and can no longer be associated with the personal data held by the issuing institution.

2.8.7. Secondary use of disclosed personal data

The data protection principles also apply to personal data disclosed in the media. Therefore, if the data are used for secondary purposes, it is important to consider the appropriate legal basis for processing them or how to inform people about the planned research. For example, when collecting data from thousands of people in social media environments, it can be complicated to ask everyone for consent or to provide information to everyone. When applying the exceptions and looking for possible alternatives, it is worth bearing in mind that the aim is to avoid harming people's interests and rights, respect people's right to decide on their data and ensure transparency and reliability of research.

In most cases, the environments from which the disclosed personal data are collected have specified in their terms of use how and for what purposes the data may be used. Some have created separate APIs which enable automatic collection of data. In all such cases, the requirements and conditions of the owner of the environment must be observed to ensure the lawful collection of data. Some companies may impose unreasonable restrictions on processing the data in their possession for research. As mentioned at the beginning of this subchapter, in the case of personal data of scientific value held by the private sector, it is not always clear whether the private interests of companies or the interests of science prevail.

Read more

- Association of Internet Researchers (AoIR) "Ethical Guidelines 3.0" (2019:14)
- AoIR's earlier guidance materials on ethics in internet research

2.9. How to respect people's rights over their data in research?

Chapter III of the GDPR sets out the data subjects' rights that they always have, irrespective of the purpose or the legal basis of the processing. Therefore, a participant in research may submit a request to the researcher about the processing of their data even if pseudonymised data from a public register are used in the research. To avoid confusion, the research team should agree beforehand on the person responsible for the personal data related to the study. That person has to respond to requests.

The data subject's rights are listed and briefly explained below.

2.9.1. Right to be informed about the processing of personal data

Articles 12, 13 and 14 provide for a general right to be informed. The controller has the obligation to draw up and publish their data protection conditions and inform the data subject of them. If it is impossible or disproportionally difficult for the researcher to contact the people, it is sufficient just to make the information publicly available. The provision of information supports the principles of transparency and fairness.

In a research study, the main source of information is the informational material given to the participant when they are asked for consent. However, it is important to consider that data subjects have the right to request information about the processing of their personal data at any time, so that the provision of the informational material or a reference to the data controller's data protection conditions may not be sufficient, and the data subject must have the possibility to contact the researcher.

Individuals retain the right to be informed about the processing of their data even if the processing is not carried out based on their consent but on another legal basis.

2.9.2. Right of access

Article 15 of the GDPR gives data subjects the right to access the data collected about them, the recipients of the data, the transfers to third countries, the sources (if the data do not originate from the data subject) and the automated decisions made based on the information. They may also ask for more general information concerning the purposes of the survey and the retention period.

When data subjects want to access their data, they must send a request to the controller, who is entitled to identify the applicant before disclosing the data to verify whether it is the same person whose data are requested. If the request is sent by email, it must be digitally signed.

Once identified, the persons have two options to obtain information about their personal data at the University of Tartu: they may come to the university to access the data or receive a copy. The only exception, in which case neither of these options is available, is if showing or providing a copy of the database or environment containing the data subject's data would harm other people – for example, if another person's data are visible. In this case, it is not possible to consult the data on the spot, and no copy can be issued. The applicant can only receive a descriptive text. It must be explained to the applicant why they cannot view or get a copy of the data.

According to subsection 6 (6) of the Personal Data Protection Act, the controller may restrict the right of access if compliance would make it impossible or significantly impede the achievement of the purpose of the research.

2.9.3. Right to rectification of data

Under Article 16 of the GDPR, the data subject has the right to demand rectification of inaccurate data and completion of incomplete data. This right is related to the principle of data quality and ensures that decisions about the person are not made based on incorrect or incomplete data.

The controller must always rectify or complete the data at the request of the data subject, except in case the controller considers the information to be complete or accurate. In the latter case, the controller must give reasons for the decision to the data subject.

Subsection 6 (6) of the Personal Data Protection Act allows the controller to restrict the right to rectify data if compliance would make it impossible or significantly impede the achievement of the purpose of the research.

Example

A few days after the interview, an interviewee contacts the researcher and asks to clarify an answer they have given. If it is possible for the researcher to do it and it is feasible at this stage of the study, the researcher should grant the request. However, if the same person repeatedly asks for clarification of one or another of their answers over a period of time, it will start to hamper the research. It is difficult to say precisely where the line between unjustified obstruction and justified clarification is drawn. Data subjects must be given reasons why they can no longer ask to clarify their responses from a specific moment.

2.9.4. Right to erasure of data

Article 17 of the GDPR provides the right to erasure, also known as the *right to be forgotten*. The erasure of data is one of the most complex rights. It must be done if one of the following circumstances mentioned in the GDPR applies:

- the purpose of processing has been fulfilled;
- the processing is unlawful;
- the person withdraws consent, and there is no other legal basis;
- the data subject objects to the processing of their data and there is no legal ground for further processing;
- erasure is necessary to comply with a legal obligation;
- the data concern the use of an information society service at a time when the data subject was a minor.

However, there are several exceptions to the right to erasure; in this case, the controller can continue processing personal data even if the data subject requests their erasure. For example, if the processing is necessary for scientific research in the public interest, the erasure of the data would make it impossible or seriously interfere with the achievement of the objectives of the study. Erasure is rather exceptional in the context of research.

Personal data can be retained for longer periods if they have been anonymised or a decision is made to retain them for archiving purposes (see 4.1). However, at the time of collecting personal data, the person must be informed of how and for how long the data will be stored.

2.9.5. Right to restriction of processing

Under Article 18, the data subject has the right to restrict the processing of personal data in four cases. In the context of research, three of them are relevant:

- If the accuracy of the data is contested, processing can be restricted for the time it takes to verify the accuracy of the data;
- If the processing of data is contested, the processing can be limited for the time to verify whether the controller's interests override those of the data subject;
- If the processing of personal data is unlawful, i.e. there is no legal basis for processing, it is possible to request the restriction of processing instead of erasure.

The GDPR lays down a few exceptions which nevertheless allow the controllers to process data with the restriction of processing: in particular, with the consent of the data subject, for legal claims, the protection of the rights of others, or reasons of substantial public interest.

According to subsection 6 (6), the controller does not have to comply fully with the right to restrict the processing of personal data if this would make it impossible or significantly hinder the achievement of the purpose of the research. The restriction of processing in the research context is very exceptional and unlikely. It can happen when a person withdraws their consent to the processing of their data but decides to request the restriction of processing instead of erasure. In such cases, the data may be retained, but their use must be limited. However, since the erasure of data mostly concerns cases where there is no legal basis for the processing, further processing would be prohibited anyway.

2.9.6. Right to data portability

Article 20 of the GDPR allows data subjects to have their data transmitted from one controller to another. This operation is subject to a few restrictions:

- Only data processed on the legal basis of consent or a contract can be transferred;
- The processing of the required data must be automated, and the transfer from one processor to another must be technically feasible;
- The transferred data must be in a structured, commonly used and machine-readable format;
- The data subject can request the transfer of only such data that the data subject has personally provided to the controller.

As most research studies are based on consent and automatic data processing, people can, in principle, always demand the transfer of their data.

Example

A person relocates to another EU member state and intends to spend the rest of their life there. Having deposited a gene sample to the biobank years ago, the person wishes to transfer all their personal data to a similar biobank in the new country of residence, so that their new doctor can obtain information about them more easily. It is only possible to transfer data the person has given to the biobank themselves, i.e. only the person's medical history, but not the genetic data generated by the biobank based on additional analyses.

2.9.7. Right to object

Article 21 of the GDPR lays down the right of the data subject to object to the processing of their data on the grounds of legitimate interest or public interest task, irrespective of the extent to which these legal grounds have been substantiated. If the objection is successful, the legal basis is cancelled, the processing becomes unlawful, and the right to request erasure or restriction arises. It is possible to request a restriction of processing while the objection is being assessed.

Based on subsection 6 (6) of the Personal Data Protection Act, the controller may restrict the right to object if compliance would make it impossible or significantly impede the achievement of the research objectives. Since objecting to the legal basis will inevitably hinder the achievement of the research objectives, it is not entirely clear what the controller should do when they receive such a request. However, it is essential to remember that if the request is not considered, the data subject may bring the matter before the Data Protection Inspectorate or the court to defend their rights.

2.9.8. Right to be protected against automated decision-making

Article 22 of the GDPR does not completely prohibit the automated processing of personal data. However, making decisions based solely on automated processing, including **profiling** (Article 4 (4)), is not permitted if such a decision produces significant effects or legal consequences for the data subject.

There are three exceptions, however. Making decisions based on automated processing is not prohibited if it is

- necessary for making or performing a contract;
- permitted by law;
- based on the data subject's explicit consent, given for automated decision-making. Such consent must be given separately from any other conditions on the consent form.

There are no exceptions to this right in the context of research. Therefore, the automated processing of personal data, including profiling, is prohibited if the decisions made during or as a result of such processing significantly impact data subjects or produce legal consequences (for example, restricting access to public services). In most cases, research does not involve making such decisions about individuals. However, in some types of applied research studies, it is theoretically possible to create and develop automated processing methods that can be later used to make decisions about individuals.

2.10. What to consider when processing the data of vulnerable people?

According to the glossary of the Code of Conduct for Research Integrity, persons or groups are vulnerable if they cannot or are unable to express their will freely (limited autonomy) or if they are susceptible to damage due to their health, work, education or other characteristics.

Consideration of vulnerability is in line with the general principles of research ethics – particularly respecting human autonomy, not doing harm and doing good. The principles of fair processing of personal data (see 2.6) require, among other things, assessing the vulnerability: the data subject's expectations must be taken into account, the exploitation and discrimination of the data subject is prohibited, and the purpose and method of processing must be ethical.

2.10.1. Vulnerable persons and groups

Children, the elderly, the pregnant, the unborn, prisoners and persons in custody, people with disabilities, ethnic minorities, the poor, the homeless, the illiterate, the unemployed and victims of violence are often considered vulnerable. The list is, of course, neither exhaustive nor universal.

Data protection does not distinguish between data subjects based on any additional characteristics; it gives everyone an equal and uniform right to the protection of their personal data. The GDPR only makes an exception for children, whose personal data protection is specifically emphasised. This does not mean, however, that vulnerability is irrelevant in data protection, especially when assessing the impact of data processing on individuals' rights and freedoms (particularly in relation to discrimination) or the ethical nature of research.

2.10.2. Vulnerable person's consent may not be voluntary

When vulnerability lies in a person's limited ability to express their will freely, it is important to determine whether and to what extent this may affect the voluntariness of the person's consent. The person's ability to understand the information and the consequences of their decision must also be taken into consideration.

Information is easier to understand when plain language or explanatory illustrations are used. Although there is no general obligation in Estonia to obtain the ethics committee's approval when involving vulnerable people in a research study, it is good practice to do so to ensure better protection of their interests and rights. Ethics committee's approval provides an additional guarantee that the vulnerability of the individuals has been adequately taken into account in the study.

2.10.3. Processing of vulnerable persons' data may jeopardise their rights and interests

The processing of personal data may involve a higher-than-usual risk of discrimination or stigmatisation of vulnerable persons. This may require a data protection impact assessment, which should include a

thorough evaluation of the possible consequences for the research subjects and propose additional measures to mitigate the risks (see also 2.14).

2.11. What to consider when processing special categories of personal data?

The processing of special categories of personal data is prohibited unless there is a legal basis for the processing and, in addition, one of the circumstances specified in Article 9 (2) of the GDPR applies. Two of them concern research most directly: if an individual has given consent to the processing of special categories of personal data or if the processing is necessary for scientific research in the public interest. In the latter case, however, there must be additional safeguards, the processing must be lawful, and the rights of the people must be respected. In Estonia, the processing of special categories of personal data must also comply with section 6 of the Personal Data Protection Act.

2.11.1. Processing special categories of personal data without consent requires the ethics committee's approval

The obligation to obtain the consent of the ethics committee is laid down in subsection 6 (4) of the Personal Data Protection Act. It applies if the research involves processing special categories of personal data without the data subject's consent. The obligation can be seen as an additional safeguard measure within Article 9 (2) j) of the GDPR. The request for consent submitted to the ethics committee must describe the measures envisaged to protect the data subject's rights and personal data.

2.11.2. Processing special categories of personal data requires additional safeguards

As special categories of personal data are highly sensitive information, they involve a risk of harming people's interests and rights. Therefore, special categories of personal data need more protection.

- Preference should be given to processing pseudonymised data or using other means of protection, such as encryption;
- Before processing, it is necessary to assess whether a data protection impact assessment (see 2.14.3) should be carried out, especially if data subjects include vulnerable persons.

2.11.3. The concept of special categories of personal data can be difficult to apply

Sometimes it is not easy to draw the line between special categories and ordinary data. For example, it may not be certain when mental health is being studied and when the mood or attitudes are.

A purpose-based approach can be helpful. When data on a person's attitudes and well-being are collected to draw conclusions about, for example, their stress level or stress tolerance, it is mental health research. However, when such data are asked to examine the person's satisfaction with a job or service, it does not concern mental health but other qualities.

2.12. How precisely should the purpose of the study be formulated?

While the purpose of the research study should be as specific as possible, it may not be known definitively at the time of data collection, and some flexibility is allowed. Therefore, the GDPR allows a broader formulation of the purpose of the study, for example, specifying the discipline or the field of research.

The purpose of the study must be distinguished from the purpose of other activities that support research. For example, if researchers know people's contact details through a consent form, they are not allowed to invite them to a new research project based on those contact details without prior consent. Similarly, the same data must not be used in the following project unless the subjects have been specifically informed. According to the interpretation of the GDPR, the secondary use of data for research purposes is allowed, but the data subject must be informed (see 2.8).

If the study has more than one purpose, and consent is the legal basis for the processing, it must be asked separately for each purpose.

2.13. When is ethics committee's approval needed?

Approval can be seen as an additional safeguard to help ensure that research is carried out ethically. In some cases, the ethics committee's approval may be a **legal obligation**; in others, the approval may be necessary to meet a funding provider's or publisher's requirements, and in some cases, for **ethical reasons**.

In most cases, asking for approval from the ethics committee depends on whether the researcher wants to conduct human research. Human research involves processing personal data and studying the participant's physical or mental health. Some human research in Estonia is specifically regulated by law, such as clinical trials on medicinal products and devices and human genetic research.

Read more
- UT Research Ethics Committee website
- European Data Protection Supervisor's preliminary opinion on data protection and research (p 12)
- Estonian Bioethics and Human Research Council's guidelines for application

2.13.1. Statutory obligation

The Personal Data Protection Act stipulates that an ethics committee assesses both research ethics and data protection standards. In Estonia, the approval of an ethics committee is required by law if

- special categories of personal data, such as health data or biometric data, are processed without the person's consent (subsection 6 (4));
- the data are processed by the biobank (section 29 of the Human Genes Research Act), including cases when the gene donor's data are de-pseudonymised (section 24 of the Human Genes Research Act);
    - personal data are issued from the health information system (section 59^4 of the Health Services Organisation Act);
- a clinical trial of a medicinal product is carried out (section 93 of the Medicinal Products Act);
- a clinical investigation of a medical device is carried out (section 21^3 of the Medical Devices Act).

In many cases, the law specifies which ethics committee to consult. The Estonian Bioethics and Human Research Council coordinates research related to the release of data from the Estonian Biobank and the Health Information System; the Ethics Committee for Medicinal Products of the Republic of Estonia Agency of Medicines coordinates the clinical trials of medicinal products; the Research Ethics Committee of the University of Tartu and the Research Ethics Committee of the National Institute for Health

Development coordinate clinical investigations of medical devices. In some cases, it is necessary to consult several ethics committees to combine data from the Health Information System and other sources.

However, in some cases, it is not specified which ethics committee should be consulted according to subsection 6 (4) of the Personal Data Protection Act. In these cases, any ethics committee recognised in the field is suitable.

2.13.2. Requirements of funders and publishers

Also, the project funder or the publishing house where the researcher plans to publish the study may require approval. For example, if the lead partner in an international project decides that approval from national ethics committees is required for the research, Estonian researchers need to obtain the approval.

Example
A medical researcher carried out a study for which he did not consider it necessary to seek the ethics committee's approval because it did not involve special categories of data. After the study was completed, the researcher wanted to publish an article on the subject in an international journal. The publisher required the researcher to obtain approval from the ethics committee. The ethics committee does not grant approvals retroactively, and so the article was not published in an international journal. Thus, while approval is not always mandatory, the lack of approval may reduce the researcher's chances to publish.

2.13.3. Ethical considerations

A researcher may also ask for the ethics committee's approval on ethical grounds, as this helps ensure that the planned research does not harm people. Consent is usually necessary when the study is carried out using a very different method from the usual one (for example, when people are misled or deceived), involves vulnerable people, or a higher risk than usual.

In addition, a researcher may always voluntarily seek the ethics committee's approval.

Read more
- UT Research Ethics Committee's guidelines and requirements
- European Data Protection Supervisor's preliminary opinion on data protection and research (p 12)

2.14. How to assess the risks associated with personal data processing?

One of the central research-ethical principles is that a research study should produce as much benefit as possible and as little harm as possible. Sometimes harm is unavoidable, but usually it can be minimised through careful risk assessment and mitigation.

2.14.1. General method of risk assessment

Research funders sometimes require researchers to assess the ethical risks associated with their study and describe how to mitigate them. When planning the study or applying for a grant, researchers may also need to assess the risks associated with their research.

Risk assessment is usually carried out in several stages.

- In the **risk identification** stage, all potential risks are listed. There is no single correct way to do it. Sometimes this task is delegated to experts or the project's lead partner; researchers involved in the study can also do it. Also rare risks must be identified, so it is always useful to involve more people.

  Even if no risks are identified, it shows the external assessors (European Commission, Estonian Research Council, Data Protection Inspectorate, ethics committee) that potential risks have been addressed, that they do not exist, and no further action is necessary.

- In the **risk analysis** stage, each identified risk is assessed based on its likelihood and potential impact. The easiest and usually acceptable solution is rating the likelihood and impact on a five-point scale ("very low", "low", "moderate", "high", and "very high") and the risk based on three colours (green, yellow and red). These are combined to construct a risk matrix table showing the probability and impact estimates. In the table below, high risk is represented by red, moderate risk by yellow and low risk by green.

  Table. Distribution of risks by likelihood and impact

LIKELIHOOD	IMPACT				
	Very low	Low	Moderate	High	Very high
Very low	0	1	2	3	4
Low	1	2	3	4	5
Moderate	2	3	4	5	6
High	3	4	5	6	7
Very high	4	5	6	7	8

  There are other risk assessment methods, and the scales of impact, likelihood and risk may differ.

- In the **risk assessment** stage, it is necessary to decide which risks are low and which are above average so that they need mitigating. Therefore, measures must be proposed to reduce the likelihood and the impact of the risk. These may be technical (secure information systems), legal (data exchange agreement with the processor) or organisational (needs-based access to data). After the risks have been mitigated, they should be reassessed until they are below average. If it is impossible to eliminate the risk fully, monitoring activities should be described to continuously monitor and mitigate the risk throughout the research. This is **risk management**, which presumes the readiness and capacity to act if one of the risks materialises or a new hazard is identified. Before managing the risks, it is necessary to agree on the responsibilities of people, on who monitors the risks, and how to respond to them.

Read more

- European Network and Information Security Agency (ENISA, 2006) report "Risk Management: Implementation principles and Inventories for Risk Management/Risk Assessment methods and tools"
- European Commission (2021) guidance "EU Grants: How to complete your ethics self-assessment", which sets out the need to assess the risks in the context of environmental damage, research safety, use of artificial intelligence, misuse of research results and processing of personal data

2.14.2. Assessment of risks associated with personal data processing

General research-ethical risks may have something in common with data protection risks[5]. For example, a research study may undermine people's right to privacy or discriminate against them. Therefore, before starting a research project, the researcher must identify the risks involved in processing personal data and the potential impact on people.

In data protection, risks should be considered in at least two cases.

First, **information security risks** must be assessed to ensure the integrity, availability, confidentiality and secure processing of personal data. This is done mainly by the university's information security specialists, who ensure that the researchers have the appropriate equipment (see 3.1).

Secondly, the **potential harm** arising to data subjects from personal data processing must be taken into account. If the study involves a high risk of harm to people's rights and freedoms, the GDPR requires the responsible person to carry out a **data protection impact assessment**, which is a specific form of risk assessment.

2.14.3. Preparing a data protection impact assessment

It is necessary to prepare a data protection impact assessment if the processing of personal data – considering the nature, scope, context and purposes of the processing – is likely to threaten the rights and freedoms of people. There is no clear and simple definition of when an impact assessment is mandatory. It is up to the controller to assess the impact of the planned processing on people.

The concept of high risk is an important element of data protection. However, the GDPR and the guidelines of the Data Protection Inspectorate define the high-risk processing of personal data somewhat differently.

1. The **GDPR** gives three examples of high-risk processing:

- systematic, extensive and automated assessment or tracking of people (including profiling), which has legal or other equivalent consequences for people;
- extensive processing of special categories of personal data or data relating to the offence;
- comprehensive surveillance of public areas.

In these cases, an impact assessment is mandatory. In other situations, the controller must assess the level of risk. The relevant factor is the harm to the rights and freedoms of people. The risk is high if there is a reasonable likelihood of harm to the rights and freedoms of individuals.

2. The **Data Protection Inspectorate** has set additional criteria for conducting a data protection impact assessment – the **scope** and **systematic nature** of the processing. According to chapter 5 of the Data Protection Inspectorate's general guidelines for data processors, data processing is systematic if it is methodical and planned. As research is by default always systematic, the controller must pay

[5] While it is common practice in IT, national defence, environmental and some other fields to consider *risk* and *oht* (threat) as different terms in Estonian (see e.g. https://akit.cyber.ee/term/52-risk and https://akit.cyber.ee/term/93-oht, https://eits.ria.ee/et/seletav-sonaraamat/o?id=96649ad7b153a7f6d3bae608d0b1cbfe, https://sonaveeb.ee/search/unif/dlall/mil/risk/1 and https://sonaveeb.ee/search/unif/dlall/mil/oht/1, https://www.riigiteataja.ee/akt/163255), they are used synonymously in this guide due to the wording of different legislative acts and guidelines.

particular attention, when assessing a high risk, to the scope of the processing, in both quantitative (a large number of data subjects) and qualitative terms (special categories of data and data on offences).

In "Making an impact assessment", the Data Protection Inspectorate points out specific cases when the scope of processing involves such a risk that a data protection impact assessment is required:

- when processing the special categories of personal data of 5,000 or more people or when processing offence data;
- when processing data posing a high risk to 10,000 or more people;
- in other cases, when processing the data of 50,000 or more people.

These figures concern the processing of personal data in Estonia. If the research study involves cross-border processing of data, the criterion of scope should be assessed on a case-by-case basis.

According to the Data Protection Inspectorate's guidelines, a high risk arises when processing, e. g.,

- data, the disclosure of which would breach the confidentiality of the message;
- people's location data in real time;
- personal data in a way that could lead to discrimination against persons with legal effect;
- personal data of children.

Data protection impact assessment

According to Article 35 of the GDPR, data protection impact assessment consists of four major parts:

- a description of the envisaged processing operations and their purposes;
- an assessment of the necessity and proportionality of the envisaged operations;
- an assessment of the risks to the rights and freedoms of people;
- the measures envisaged to mitigate the risk.

When preparing the impact assessment, other documents relevant to the personal data processing and directly related to the study should be consulted, such as research records reflecting the data processing method, policies on granting access rights, contracts, etc. If necessary, the specialist for data protection should be involved. In the case of international research, the organiser may require an impact assessment to be carried out under the organiser's rules in the country of research.

The specialist for data protection must be involved when the impact assessment has found that a major risk persists and the proposed measures do not entirely eliminate or sufficiently mitigate it. If necessary, the options to mitigate the risks are explored in cooperation, in consultation with the Data Protection Inspectorate.

Read more

- Data Protection Inspectorate's general guidelines for data processors, chapter 5 "Making an impact assessment" and annex 1 Checklist for making an impact assessment"
- European Network and Information Security Agency (ENISA, 2006) report "Risk Management: Implementation principles and Inventories for Risk Management/Risk Assessment methods and tools"
- Data Protection Inspectorate's sample impact assessment

2.15. What to consider when processing children's personal data?

Under the GDPR, children's personal data require a higher level of protection than usual. Minors have exactly the same rights over their personal data as adults, but they cannot give consent to the processing of their personal data due to incapacity. Both the UN Convention on the Rights of the Child and the Child Protection Act require researchers to always set the child's best interests as a primary consideration in decisions concerning children. Therefore, account must always be taken of what children want or prefer, even if their parents or guardians have given their consent for the child.

2.15.1. Minors cannot give consent but must be asked to assent to the processing of their data

The legal framework for personal data protection presupposes that the signatory of a contract or consent giver has the power of representation. Estonian law considers that a person aged 18 years or over has legal capacity, so the consent of the parent or other legal representative is required in the case of a minor. However, the law allows the legal capacity of minors to be extended: for example, under section 8 of the Data Protection Act, children aged 13 years or over have the right to give their consent to the use of information society services (e.g. social media). No such exception has been granted for research, and using IT solutions to collect personal data does not make a study an information society service.

However, under the GDPR, the rights of minors must be considered: minors should be informed about the processing of their personal data and their free will must be respected. In the case of research, it is good practice to ask minors for their assent to the study, even if this is not consent in legal terms. If the minor does not assent to the processing of their personal data, involving the person in the study is not allowed, even if the parent has given consent. Unintentional involvement is contrary to the code of conduct for research integrity and the principle of fair processing of personal data. The same applies where a minor wants to withdraw their assent but the parent has not withdrawn their consent – fair processing and respect for the child's autonomy require the researcher to take account of the child's wishes.

2.15.2. Children must be informed about the use of their data in plain and clear language

In the case of consent-based studies, information should also be provided to children for whom a separate information form may be drawn up. It must be in age-appropriate wording or in the form of images, icons or animations. At the same time, children should be given an easy opportunity to ask questions.

Read more:

Irish Data Protection Commission, 2021 report "Children Front and Centre: Fundamentals for a Child-Oriented Approach to Data Processing", which provides recommendations for taking into account the best interests of the child and communicating information to children

2.15.3. Legitimate interest cannot be the legal basis for processing a child's personal data

Although the legitimate interest is exceptional in research, the GDPR does not explicitly prohibit processing children's personal data based on legitimate interest. However, this requires weighing the best interests and rights of the child. The emphasis of the GDPR implies that the rights and interests of the child take precedence over the normal consideration of legitimate interest.

2.16. What to consider when processing the data of deceased persons?

The death of a person is a complex issue from a data protection point of view, as neither the GDPR, the explanatory memorandum to the Data Protection Act, the guidelines of the European Data Protection Board, nor those of the Data Protection Inspectorate address what happens to personal data after the person dies.

The GDPR only protects the data of living people, so member states can restrict the processing of the data of the deceased.

Section 9 of the Data Protection Act specifies the conditions for processing the data of deceased persons. The data of deceased adults are protected for ten years, and the data of deceased minors for 20 years after death. During this period, all data protection principles must be complied with, including finding an appropriate legal basis for the processing. After this period, the data will no longer be subject to data protection and will have to be used in accordance with general principles of research ethics and professional practices.

2.16.1. Protection of deceased persons' data serves to protect other people

The data of a deceased person may also concern their friends and relatives. It is, therefore, important to remember that the focus of protection is not on the interests and rights of the deceased person but on those of their friends and relatives. If the researcher has access to the deceased person's files, documents, diaries or other materials, the living persons' data included in them are still protected.

For example, it may be difficult to process data on the circumstances of death if the death was caused by a person who is still alive or if the cause of death can be used to draw conclusions about the health of other people. There may also be restrictions on using the information on other bases, for example, where it may reveal the adoption secrecy within the meaning of section 164 of the Family Law Act.

2.16.2. After death, the right to give and withdraw consent passes to successors

In the case of consent-based research, it should be borne in mind that the successor can give consent to the processing of the deceased person's data. If the data are to be collected during the ten-year period of protection after death, this will mean an additional burden for the researcher, who will have to identify and contact the successors. In the case of multiple successors, consent can be given and withdrawn by any of them. Under subsection 18 (1) of the Law of Succession Act, in the absence of successors, inheritable rights are transferred to the local government of the data subject's place of residence, from which consent can be sought. If it is known in advance that it will be difficult to obtain consent for collecting the data of the deceased person, consideration may be given to whether another legal basis is available.

2.16.3. Other rights of the data subject are not transferred to successors

A somewhat more complex question concerns the extent to which the successors can exercise the rights of the data subject listed in Article 3 of the GDPR. Since the withdrawal of consent may render the processing of the personal data of the deceased unlawful, it could, in turn, follow that the successor could also submit a request for erasure under Article 17 of the GDPR. At the same time, the explanatory memorandum of the Personal Data Protection Act does not mention whether and to what extent other

rights of the data subject concerning their data can be inherited. However, if the successor were to request the erasure of the data, it would be possible to rely on the general exceptions mentioned in clause (d) of Article 17 (3) of the GDPR, namely archiving in the public interest, scientific research or historical research (see also 2.9.4).

2.16.4. Researcher is not obliged to keep an account of the life and death of the subjects

Article 11 of the GDPR states that the controller is not obliged to collect additional information in order to identify the data subject for the sole purpose of complying with the GDPR. Although Article 11 does not explicitly refer to identifying whether the data subject is alive, it can be assumed that a similar principle applies, especially since the GDPR does not protect the data of deceased persons. Thus, it can be presumed that a researcher is not obliged to keep an account of who of their subjects is alive and who is dead, only to identify who is entitled to withdraw consent. Moreover, such continuous monitoring of being alive could be seen as a separate objective, which has nothing in common with the objectives of the research and would therefore require a separate justification.

Inevitably, it can happen that, after the data subject's death, the subject's friends and family do not know about the person's participation in the study, and the researchers do not know that the data subject has died. Such situations need to be approached case by case. If the data subject has given consent before death, the research can continue, but if a successor wishes to withdraw consent, this must be taken into account. In this case, the successor must prove that the data subject is deceased and he or she is the successor. A death certificate is a suitable means of proof, but the researcher does not need to retain it anywhere.

There are no clear practices for taking into account the possible death of the data subject. If it is known at the time of planning the research that the death of the data subjects is likely – for example, if very old or terminally ill people are studied or if the subjects do very dangerous work – solutions can be considered for communicating the research information to their families at an early stage.

2.16.5. Data about the deceased may be processed on other legal bases

If the research is carried out on a legal basis other than consent, the death of the data subject does not make much difference to data processing. The 10-year protection of deceased persons' data (20 years for minors) is only valid with the consent of the data subject, which can be given and withdrawn by the successor after death. Where the legal basis is, for example, a task in the public interest, further processing can be carried out without the consent of the data subject and the successor (see also 2.4). Subsection 9 (4) of the Data Protection Act also exceptionally allows the processing of the name, sex, date of birth and death, the fact of death, and the time and place of burial of a person without the consent of the successor.

3. Doing research: data collection and analysis

This chapter gives an overview of the data protection issues that may arise in the course of research once the processing of personal data has already started. The chapter deals with ensuring security, anonymisation, pseudonymisation and responding to possible breaches.

Despite careful planning, unexpected problems may arise during the research process. It is also important to be aware of the possibility that people involved in research may wish to exercise their rights and make requests or other demands for their data.

3.1. How to ensure the security of personal data processing?

For personal data, **security** means, in particular, ensuring the integrity, availability and confidentiality of the data. Security is ensured by both technical means (e.g. equipment, software) and organisational measures (e.g. access rights, training). To ensure that security is maintained, the adequacy of the tools and measures used must be reassessed from time to time. For example, the measures taken for a research project conducted five years ago may no longer be adequate for a new project.

The **integrity** of personal data is compromised by any activity that involves unauthorised modification or deletion of data, such as theft, cyber-attacks, technical failures of equipment and systems, or accidents. In the case of research, it can even mean the failure of the study because there is no longer enough data or the data cannot be analysed. Integrity can also be undermined by negligence when researchers accidentally alter or delete data. Back-up and data processing software that does not modify the underlying data during analysis can help protect against such errors.

Availability requires that personal data are easy to use for the purpose for which they were collected. For example, storing data on an offline device may be secure, but it can significantly reduce availability if researchers have to physically go somewhere to analyse the data each time. At the same time, the most convenient and popular tools are not always the most secure. So it is important to find the right balance.

The **confidentiality** of personal data is compromised when private information becomes known to unintended outsiders. For example, when leaving your workplace without locking your computer or office or working on a laptop or phone in a public place (public transport, café, park), nearby people may see files containing personal data. In particular, organisational measures can help prevent such problems. Malicious attacks aimed at stealing or leaking personal data are an even bigger threat. It is, therefore, a good idea to pseudonymise personal data to reduce the damage to people's privacy caused by possible data leakage or theft.

Below, we have commented on some of the advice given in the European Data Protection Board's Guidelines 4/2019 on Article 25 of the GDPR, "Data protection by design and by default", and explained how to comply with them in the case of research.

3.1.1. Systematic management of information security

Systematic information security risk assessment and the implementation, monitoring and improvement of security measures are primarily carried out by the University of Tartu. The university is also responsible for ensuring that the information systems, tools and services provided to researchers are

sufficiently secure to process personal data. Systematic approach also implies assessing and managing data protection risks (see also 2.14).

It is the researcher's responsibility to be aware of information security risks, follow agreements and guidelines, and seek assistance when necessary. The university's guidelines on cybersecurity may be helpful.

3.1.2. Needs-based access to personal data

Access rights management is one of the most common security measures a controller may implement. A prerequisite for restricting access is a clear overview of the researchers who need to process personal data for research purposes. It is important to ensure that those who do not need to process personal data cannot do so intentionally or unintentionally. If students are involved at any stage of the research, a confidentiality agreement must be concluded with them.

It may be necessary to retain log files to verify access rights, especially in the case of long-term research where large amounts of sensitive data are processed. It is also worth paying more attention to access rights if it is known that members of the research team will change more frequently than usual.

3.1.3. Secure transfer of data

If personal data need to be transferred to another researcher or research institution, it must be ensured that their integrity and confidentiality are not compromised in the process. For example, where possible, the transfer of a copy of personal data by email should be avoided if the recipient can be given access through the information system where the data are stored. If sending the data by email is the only possible solution, the data should be encrypted, or other measures should be taken to avoid the possibility that the data can be seen by anyone other than the addressee.

An example of a security risk is transferring personal data via a memory stick or other external data carrier that could be lost. However, if this is done, both the data carrier and the data file on it should be encrypted to ensure security.

3.1.4. Secure storage of data

When storing data, they must be protected against unauthorised modification and access. This depends on the opportunities available for the researcher and the tools used:

- only the university's work computer can be used to store personal data collected for the university's research. This is usually a laptop the researcher takes to work, on a business trip and home. A situation where unauthorised persons could access a work computer containing sensitive data must be avoided;
- the university cannot be held responsible for data processing on the researcher's personal computer. If data becomes public when using a private computer (a data breach), the researcher is liable to the data subjects and the university; must remedy the situation for the data subjects and report to the Data Protection Inspectorate;
- a device not connected to the network is more secure than a connected device because it is much more difficult to attack;
- a single-user device is more secure than a multi-user device.

There are many other criteria to be taken into account for securely storing personal data, such as the sensitivity of the data, the amount of data, the availability of the data, the possibility of managing access, and the equipment and software used for processing.

3.1.5. Backing up data

Back-ups help ensure the integrity and availability of data if they are destroyed or significantly damaged by accident, malicious activity or negligence. The 3-2-1 rule is used in data management: data should be backed up in at least three copies, on at least two different data carriers or environments, one of which should be located elsewhere.

One of the three copies should be the working copy, where data can be modified, supplemented and deleted during the work. The second copy is necessary in case of damage to the working file, accidental deletion of important data or destruction of the working file. The third copy is a backed-up copy stored on another device or environment (cloud, etc.) and is not easily accessible. For example, it is good to store data not only on a work device but also on a university's network drive, server or cloud environment. When using a cloud service, it is important to ensure that the university has a contract with that environment. Storing data on two data carriers helps ensure that if something happens to one (fire, flood, theft, etc.), the data will remain available on the other. In this case, it is important to assess the risks: for example, having a back-up in another room of the same building may not protect against fire.

At the same time, back-up must be well thought out, fit for purpose and in line with the data minimisation principle. Data cannot be duplicated just in case (without a clear purpose and need). For advice and assistance on backing up, contact arvutiabi@ut.ee.

Read more:

Chapter "Storage and back-up" of the guidelines on creating the data management plan by the University of Tartu Library

3.1.6. Awareness of the possibility of breaches

A breach is the misuse of personal data (disclosure, deletion, mistakes in asking for consent, etc.). Awareness of the main data protection principles and potential risks in research helps to prevent breaches. Prevention requires far less resources than dealing with the consequences of a breach.

Breaches should be reported immediately to the senior specialist for data protection by email to andmekaitse@ut.ee (see also 3.5).

3.1.7. Appropriate services, software and tools for processing personal data

The tools used to process personal data must ensure the secure processing, confidentiality, availability and integrity of personal data, as well as the legal protection of the data subject. A distinction can therefore be made between tools based on whether the data are only accessible to the processor or also to the creator of the tool and the service provider, e.g. the owner of the survey environment, the repository administrator or the company licensing the software.

If the data move outside the university, the tool's suitability for research purposes must be carefully assessed. For that, consult the data protection policy of the service provider or software owner. If the

processing of data is not described in sufficient detail or the policy raises doubts, the service or software is probably not reliable.

In case of questions or doubts, consult the university's chief information security officer. It is sometimes possible to mitigate legal and technical risks in a contract with the service provider, for example, by agreeing that data will only be stored on the university's servers.

The services and software provided can work in three ways.

- **They do not transfer any personal data**: such solutions are always more secure, as the data being processed remain only on one device or information system the researcher uses. For example, the software for qualitative data analysis usually stores interview transcripts on the researcher's device and does not transfer the data anywhere. In this case, security depends on the researcher's actions, including where and how the person stores the interview recordings, transcripts or parts of them. It should be noted that project files created by the software may contain personal data.
- **They transfer personal data within the research institution**: an example of such a solution is a cloud service managed by the university or software licensed from a company that ensures that the data are only stored on the university's systems. It is important to remember that the researcher must verify the solution's security.
- **They transfer personal data outside the research institution**: in this case, adequate security and legal protection must be ensured. Particular attention should be paid to solutions where personal data are automatically transferred outside the EU, for example, where all the data entered are stored on servers located in third countries (see 3.2). In such cases, an additional safeguard, such as a contract between the university and the service provider, is generally required. If the data are stored on a server in an EU member state, this offers adequate legal protection, but care must be taken to ensure that this is done securely.

3.2. What to consider when personal data is transferred from one country to another?

The general principle applies that data flows between countries must be subject to adequate data protection in the country of destination. For some countries, the protection is considered adequate; for others, the University of Tartu as the data controller must take additional measures. Therefore, if you wish to transfer personal data to a third country, always consult the university's senior specialist for data protection.

3.2.1. European Union member states, Iceland, Liechtenstein and Norway

If personal data are transferred to countries in the European Economic Area (EEA), the GDPR ensures adequate protection, and no additional restrictions or requirements apply. The general principles must be respected: processing must be lawful, fair, transparent, secure, purposeful and minimal. There must also be a contract for the data transfer.

However, the laws of the different EU member states on research and research ethics differ somewhat. Therefore, it is advisable to discuss with partners what the requirements are for personal data in the other country.

3.2.2. Third countries with an adequate level of data protection

If data are transferred to a third country outside the EEA (a third country), the level of data protection there must be assessed. The European Commission has found that the level of data protection is adequate in Andorra, Argentina, Canada, the Faroe Islands, Israel, Japan, the Republic of Korea, Switzerland, Uruguay, New Zealand, the United Kingdom and the British Crown Dependencies of Guernsey, the Isle of Man and Jersey. There, no additional safeguards are necessary, and the same requirements apply as to EU member states.

Read more:

List of countries with an adequate level of data protection on the website of the European Commission

3.2.3. Other third countries

In 2016, the European Commission assessed the EU-US data protection framework Privacy Shield as adequate in terms of the level of protection, but this assessment was invalidated by the Court of Justice of the European Union in a 2020 ruling – so the level of data protection in the US is currently not adequate. Therefore, the exchange of personal data between Estonia and the US requires additional safeguards. This could be done, for example, through a data transfer agreement with the institution in the US or any other measure mentioned in Article 46 of the GDPR. In addition, the transfer of personal data to the US is subject to a data protection impact assessment (see 2.14.3).

For all other third countries whose level of data protection has not been recognised as adequate by the European Commission, the controller will also need to implement additional safeguards, which in most cases implies the conclusion of a separate agreement with a cooperation partner in the third country.

3.3. Why and how to pseudonymise personal data?

According to Article 4 of the GDPR, pseudonymisation means the "processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person". Thus, by pseudonymisation, all the data or identifiers that allow a person's direct or indirect identification are replaced by a pseudonym, after which the person is no longer identifiable.

However, pseudonymisation is reversible. The additional information referred to in the GDPR, such as a key, a code or any other identifying information, can be used to re-establish the original link between the data and the person. In its simplest form, this additional information can be, for example, a table of identifiable data and the pseudonyms assigned to replace them. To ensure security, this information must be carefully protected.

Thus, *pseudonymisation* is a generic term for all data processing methods that allow both de-identification and re-identification of a person. It should not be forgotten that pseudonymised data are still personal data. Even if researchers cannot identify an individual by looking at the data, the data must be treated the same way as identifiable data, and all data protection principles must be respected.

There are different understandings of pseudonymised data in Estonian law. For example, the Data Protection Inspectorate has drawn attention to the need to amend section 7 of the Human Genes Research Act. It reads: "The provisions regulating the processing of personal data do not apply to the processing of pseudonymised tissue samples, pseudonymised descriptions of DNA and pseudonymised descriptions of state of health if such tissue samples, descriptions of DNA and descriptions of state of health are processed as a set of data and on the condition that the set of data to be processed contains DNA samples, descriptions of DNA or descriptions of state of health of at least five gene donors at a time." However, Recital 26 of the GDPR states that pseudonymised personal data should be considered information on an identifiable natural person. Thus, pseudonymised genetic or health data cannot be classified as non-personal data to which neither the GDPR nor the Personal Data Protection Act apply.

3.3.1. Causes and timing of data pseudonymisation

According to the GDPR, pseudonymisation enhances the security of the processing of personal data and data protection by design. The principle of minimisation must be respected: if the processing does not require the identification of the data subject, the processing of personalised data is not justified. Thus, pseudonymisation does not only concern the transmission of personal data but also the work of a research institution or a research project to reduce the number of researchers who can identify individuals based on the data.

The more sensitive the data, the more necessary pseudonymisation may be. It should also be considered when data are transferred to third parties.

Personal data should be pseudonymised as soon as possible. For example, in a research project with several partners abroad, this should be done immediately after data collection and before starting the analysis or transferring the data to project partners.

3.3.2. Pseudonymisation entities

As stated in the 2019 guidelines "Pseudonymisation techniques and best practices" by the European Union Agency for Cybersecurity (ENISA), a pseudonymisation entity can be either a data controller, a data processor or a trusted third party. However, the responsibility for the security of data processing always rests with the controller.

Pseudonymisation is certainly necessary in the case of a joint study between several institutions. For example, two or more partners may be joint controllers, but they must agree that the research institution collecting the personal data pseudonymises the data before transferring them to the joint controllers. In this way, the principles of minimisation and security in data processing are respected. Similarly, the personal data may be pseudonymised by the processor (e.g. the survey company) before transferring the data to the research institution.

3.3.3. Methods of data pseudonymisation

When setting up an institution- or project-based pseudonymisation policy, the ENISA guidelines on pseudonymisation techniques and best practices, which recommend a risk-based approach to the choice of pseudonymisation method, can be used. The risks considered include potential attacks on pseudonymised datasets, the sensitivity of the data, the availability of the data and the need to protect the data.

In most cases, pseudonymisation is not merely replacing a person's name and personal identification code with a pseudonym but needs to consider all data that can be easily associated with that person.

The type of data is also important – for example, pseudonymisation of identifiers is not appropriate for images and pictorial data, but it is useful if the file names of images or metadata of images contain identifiers that may allow the identification of individuals. The processor of pseudonymised data must be unable to identify the persons behind the data.

There are different ways to replace identifiers with pseudonyms:

- **Counter** uses numbers generated based on a predefined sequence to replace identifiers. The advantage of this method is simplicity and the fact that the number assigned by the counter has no direct relationship to the identifier to be replaced;
- **Random number generator** also replaces identifiers with numbers but with random ones. Random numbers are safer than the counter because pseudonyms are not generated sequentially. However, the disadvantage is that two identifiers can be associated with the same pseudonym. The likelihood of that can be reduced by generating longer numbers;
- **Cryptographic hash function** allows arbitrary-length identifiers to be encrypted into a fixed-length code. The hash function is a one-way method, meaning it is extremely difficult to compute the original value from the hash code. It is also collision-free, i.e., no two identifiers result in the same hash code. However, since the same input results in the same hash, it is possible to depseudonymise the data by knowing the original identifier and the hash function;
- **Message authentication code** (keyed hash function) uses a secret key to generate the pseudonym in addition to the hash function. It provides the additional assurance that it is not possible to compute the identifier from the hash code;
- **Symmetric encryption** uses a single secret key for encryption and decryption;
- For smaller datasets, pseudonyms can also be created **manually**, e.g. by replacing a person's identifiers and quasi-identifiers[6] with the generic name *interviewee A* or *subject M45*. However, replacing names with initials or pseudonymising some quasi-identifiers may not provide very strong protection against identification. Random pseudonyms are almost always safer than systemic ones.

Regardless of the method, it is important to protect the pseudonymisation secret – the key, code, method or other data that allow the pseudonym to be associated with an individual. If this secret is leaked due to a cyberattack, it is possible to identify people in all the datasets compiled in the pseudonymised form. Such an attack is even more dangerous if the same pseudonymisation method is used all the time. In this case, a very serious privacy violation can occur.

To maintain secrecy, as few people as possible should have access to the depseudonymisation information, but it is good to have more than one such person in case something happens to the owner of the information or if the person leaves the job.

[6]A quasi-identifier is gender, age, nationality or any other characteristic that, on its own, cannot uniquely identify a person, but can be used in combination with other characteristics to create a direct identifier that refers to a specific person.

3.4. Why and how to anonymise personal data?

According to the Opinion 05/2014 on Anonymisation Techniques of the European Data Protection Working Party, anonymisation is the processing of data in an irreversible way, i.e. after which it is no longer possible to identify individuals by any reasonable and likely method. As a result, anonymised data are not vulnerable to attacks: even if all the data fell into the hands of an attacker, they could not be personalised. Therefore, anonymised data are not considered personal data.

In its Guidelines 04/2020 on the use of location data and contact tracing tools in the context of the COVID-19 outbreak, the European Data Protection Board has stated that only entire datasets can be anonymised, not single data patterns. From a legal perspective, it is unclear to what level the dataset must be processed to be considered anonymous. Anonymisation methods offer varying degrees of protection and often depend on the specific dataset.

3.4.1. Causes and timing of data anonymisation

The anonymisation of personal data helps to protect people's privacy and supports the principle of minimisation: if research objectives can be achieved with anonymised data, anonymisation should be preferred in all cases.

As anonymised data are no longer considered personal data, they can be used and shared more freely. They can be forwarded to partners in a research project, stored as open data in repositories or sent to other persons and institutions interested in them.

Anonymised data also make it easier to ensure the security of data processing. The only risk to bear in mind and assess from time to time is the possibility that individuals in anonymised datasets may become re-identifiable as technology evolves and new datasets are added.

Anonymisation almost always reduces the availability of the data. If the data are voluminous, multivariate or qualitative, anonymisation may prevent their use or render them useless by distorting the data. For example, anonymising qualitative data from social sciences (interview transcripts, texts) may reduce the possibility of reusing them. Moreover, anonymised material does not allow the replication of scientific analyses based on personal data.

Data can also be collected anonymously from the start, but if unique identifiers are stored in the process (e.g. computer IP address), post-processing is necessary to exclude the possibility of indirect identification of individuals. Therefore, it is important to carefully assess whether the planned method allows for collecting the data anonymously from the start or whether it is necessary to anonymise the data after the data collection or the completion of the study.

3.4.2. Anonymisation entities

The University of Tartu is responsible for anonymising personal data, but the university researcher who has the necessary knowledge, skills and resources is responsible for the specific anonymisation activities. Anonymisation may also be carried out by persons not directly involved in the research, provided that the data subjects have been informed of that in advance and that the lawfulness and compliance with data protection principles of such anonymisation are ensured.

Where secondary data are used, they may be anonymised by the institution issuing the data.

3.4.3. Methods of data anonymisation

The means of anonymisation largely depend on the nature and amount of personal data. Therefore, it is necessary to assess to what extent the chosen method prevents the association of the data with the person and whether this result is irreversible.

The three most common methods of data anonymisation:

- **Removal** involves deleting or permanently replacing all directly identifiable features (name, personal identification code). Removing direct identifiers does not immediately guarantee anonymity, as a person can also be identified from other data: for example, they can be distinguished by a unique combination of identifiers or when different datasets are combined;
- **Randomisation** implies the random distortion of data based on certain values or characteristics. As the data gets distorted, randomisation may not be suitable for the publication of scientific data. On the other hand, randomisation is used to protect large public datasets against re-identification;
- **Generalisation** involves grouping values by characteristics. For example, birth years can be grouped into age ranges, wage amounts into wage ranges, etc. Generalisation helps to ensure that an individual is not identifiable but has the disadvantage of reducing the degree of precision of the value.

In addition, depending on the data to be anonymised, some specific cases can be distinguished.

- **Anonymisation of an extract of a dataset**

  As anonymisation must be irreversible, there must be no copy of the original data that could be recombined with the anonymised dataset. However, it is possible to make anonymised extracts of the dataset for public disclosure so that the original data are preserved. Once an extract has been made, it must no longer be possible to link it to the original data.

- **Anonymisation of pseudonymised data**

  If previously pseudonymised data are anonymised, the secret key must be deleted. In addition, the adequacy of the pseudonymisation should be assessed: if only the direct identifiers are replaced by the pseudonym and not the data values, the dataset may contain unique quasi-indicator combinations that facilitate the identification of individuals. In this case, in addition to deleting the key, the data should be further processed – e.g. generalised – to exclude the possibility of indirect identification. However, if the data have been correctly pseudonymised, permanent deletion of the key may be sufficient.

To increase transparency, the method of anonymisation should be precisely described to the data owner so that they can assess whether and to what extent they consider such processing to be adequate. This is particularly necessary when anonymised data are published as open scientific data.

3.4.4. Avoiding the linking of data and persons

To reduce the possibility of attributing data to an individual, it is necessary to look at the characteristics of the dataset, such as the structure, type or amount of data. For example, surveys with a very narrow sample, which collect very precise values for many social characteristics or contain voluminous free-text responses, reduce anonymity. The European Data Protection Board's Guidelines 04/2020 on the use of location data and contact tracing tools in the context of the COVID-19 outbreak addresses cases where

data can be linked to an individual after anonymisation. To avoid this, it is important to be aware of the weaknesses of anonymisation.

- The possibility of **singling out an individual** arises when the anonymised dataset contains unique identifiers, such as IP address, device ID or a combination of quasi-identifiers. In the latter case, however, additional steps are needed to identify the individual, as several datasets for the same person need to be merged.

  Example

  If the dataset has only one entry about a person who is male, aged between 31 and 40, has a higher education, works in sub-unit Y of institution X and has ten years of experience, he is identifiable as an individual. In such a case, he could be identified merely based on a public list of staff members of institution X, together with their photos and brief CVs. He is also likely to be identifiable by all the employees of the same institution.

  The main method to avoid the identification of an individual is **k-anonymity**, which requires that for each combination of quasi-identifiers, there are at least k different matches in the dataset. The value of k-anonymity has to be chosen by the researchers themselves, depending on the sensitivity of the data and the specificities of the dataset.

- The **possibility of linkability** arises when two datasets can be matched based on some characteristics (e.g. the same quasi-identifiers). In such a case, linking two datasets may reveal that they both contain a similar unique combination of quasi-identifiers, which allows to obtain additional information about some individuals and to identify them. Merging the datasets has been the main way in which data that were initially considered anonymous have nevertheless been used to identify individuals.

Read more:

- Linking genealogical databases and anonymous DNA donor data: Bohannon, J. (2013). Genealogy Databases Enable Naming of Anonymous DNA Donors. *Science*, 339(6117), 262
- Identification of Netflix users based on movie ratings data thought to be anonymous: Narayanan, A.; Shmatikov, V. (2008). Robust De-anonymisation of Large Sparse Datasets. *IEEE Symposium on Security and Privacy,* 111–125

- **Inference** is possible if additional information is known about the person in the dataset. For example, people who work or study together know more about each other and can recognise each other from datasets without direct identifiers. Additional information may simply be the knowledge that a person you know took part in the survey – hence one of the data lines is about them. It is also possible to recognise a person by their voice or by the use of words characteristic of them. A special case of inference is when a person recognises him or herself from the data.

  It is quite difficult to avoid inference, as the amount of possible background knowledge is indefinite and depends on the individual. It should also be kept in mind that k-anonymity may not protect against inferred knowledge if the protected characteristics are homogeneous.

  Example

  The dataset has at least five (k = 5) matches for the combination of four characteristics: female, 30–40 years old, from Tartu, employment status: on parental leave. One needs to know only three of the characteristics to obtain additional information on the fourth characteristic or to identify

the person. In such a case, the **l-diversity** indicator should be considered, which assumes that there are also different values for each sensitive characteristic. For example, l-diversity = 2 would assume that for these five 30–40-year-old women from Tartu, the employment status should have at least two values: some on parental leave, some actively employed, unemployed, etc.

- At some point, due to **advances in technology or merging with new datasets**, it may become possible to identify anonymised individuals, especially if the data are stored for decades. In this case, the risk of identification must be assessed, and it must be taken into account that if the data become identifiable, the data protection principles will apply again. The data controller must then assess reasonable identifiability and demonstrate that the data can indeed be considered anonymous.

3.4.5. How to conduct an anonymous survey?

An anonymous survey collects responses in such a form and manner that respondents cannot be identified in any way.

When collecting data from people in an online survey, it should be borne in mind that IP addresses are also personal data (see also 1.3.2) and, when stored, may render the individuals identifiable. In this case, the survey is not anonymous but collects personal data. However, anonymisation is possible if the data are post-processed – for example, if IP addresses are permanently deleted after the data collection. Participants in the survey must be clearly informed of both the collection of personal data and their subsequent anonymisation.

Some survey environments also allow you to configure what additional data are collected and stored in the survey. If it is possible to turn off the collection of IP addresses and other data, the data collection can be considered anonymous. However, it is important to be aware of the possibility that even very carefully configured survey responses can make a person identifiable – for example, by asking for contact details.

At the university, using the LimeSurvey environment is recommended, which offers additional options to ensure anonymity, incl. turning off the automatic recording of the respondent's IP address. If the researcher uses LimeSurvey or another environment recognised by the university, support is available from the IT helpdesk (arvutiabi@ut.ee) if questions arise. When using environments not recognised by the university, the IT helpdesk cannot assist the researcher in case of problems.

3.5. What to do in the event of a data breach?

According to Article 4 (12) of the GDPR, a **personal data breach** means a breach of security leading to the accidental or unlawful destruction, loss, alteration, unauthorised disclosure of, or access to, personal data transmitted, stored or otherwise processed.

Examples of a data breach:

- Data which should not be public have become public;
- Data have been accidentally deleted or are no longer accessible for the necessary operations, even if a back-up copy is restored;
- Unauthorised persons have accessed the data: for example, students are involved in the data analysis phase of the study, but no confidentiality agreement is signed with them beforehand;

- Written consent is sought for the study to process the data for a specific purpose, but the data are used for purposes unrelated to the study;
- Personal data are collected based on an opt-out mechanism, so the data subject must take steps to refuse the data collection.

A breach may harm an individual and their interests by causing physical, material or non-material damage. To avoid this, the university, as the data controller, must have a full overview and control over data processing.

3.5.1. Data breaches must be reported immediately

If a personal data breach occurs at the university, the university's senior specialist for data protection (andmekaitse@ut.ee) must be informed immediately. Action should be taken as soon as possible, for example, to stop unauthorised access to or misuse of the data or other breaches. To prevent similar incidents in the future, it may also be necessary to report the incident to the IT helpdesk (arvutiabi@ut.ee).

The senior specialist for data protection must also be informed if there is only a suspicion of a breach – this will help to clarify the circumstances.

3.5.2. Be prepared to share information after a breach has been reported

Under Article 33 (5) of the GDPR, the university must document any personal data breaches, including the facts relating to the data breach, its effects and the remedial action taken. Therefore, an investigation will be launched after a breach, and if necessary, additional information will be gathered. People involved in the breach should be prepared to provide written explanations or the required materials to the data protection specialist.

Time should be taken to address the causes and consequences of the breach. It is very important to resolve the situation that has arisen (stop the data leak, inform the data subjects, assess what happened and why, determine how many people the data were disclosed to, review the whole process, etc.). This is all very time-consuming.

The university's senior specialist for data protection will also inform the Data Protection Inspectorate about the breach, which in turn may open infringement proceedings against the university.

3.5.3. Possible consequences of the breach

Once the investigation into the breach is complete, solutions must be found to ensure that a similar incident does not happen again. These may include implementing additional protection measures, raising awareness, adjusting procedures, etc.

Under Article 82 (1) of the GDPR, anyone who has suffered material or non-material damage due to a personal data breach has the right to receive compensation from the controller or processor for the damage suffered. Chapter 6 of the Personal Data Protection Act lists the amounts of the fines that apply in case of a breach of the controller's obligations. The Data Protection Inspectorate may impose a non-compliance levy upon failure to comply with a precept. The university may also hold an employee liable if it is found that the breach was due to the employee's negligence.

In addition to the Personal Data Protection Act and the GDPR, sanctions for personal data breaches are also provided for in sections 157–157^2 of the Penal Code. The Penal Code allows for the prosecution of

the natural person who committed the offence, i.e. the specific university employee who is at fault for the breach.

Read more:
- University's data protection guidelines
- University's guidelines on cybersecurity
- Clauses 55–57 of chapter IX of the university's Documentary procedure rules on the action in the case of non-compliance

3.6. What to do if the data subject makes a request about their data?

If a person contacts a researcher or the university with a request about the processing of their data, a response must be provided within 30 days.

- The **request** must be documented. The University of Tartu, as the data controller, has a document management system as its main documentation tool, where the recipient must upload the request. Based on the documentation, a deadline for the response can be set, which makes it possible to check whether each request has been replied to. Other documentation systems may be used to help the university keep track of incoming requests and respond to them, but the request must still be registered in the document management system.
- **The data subject must be identified.** As personal data must not be disclosed to third parties, it must be established that the person who made the request is indeed the data subject whose data the request concerns. To identify the requester, the person must send a digitally signed request before the data can be provided.
- **Before responding, the feasibility of responding to the request must be established.** The request must be based on a right of the data subject (see 2.9). For example, if an overview of the personal data being processed in the research is requested, it must be provided to the person filing the request. Some of these rights – such as the right to object – only apply in specific cases, for which assistance can be sought from the data protection specialist. If the request includes a request for deletion of the data, the feasibility of the request must also consider the possibility that the personal data may be contained in back-ups, from where they may be difficult to delete.

  The most complex request concerns the right of access to one's data. The university must comply with reasonable requests, i.e. no data subject's right is absolute. If it is not feasible to respond to a request, the reason for this must be explained (see also 2.9.2).

Regardless of whether it is feasible or not to respond to the data subject's request, the data subject must in any case be given a response within the time limit. There is no standard reply form or procedure, so the form of the reply will mostly depend on the question.

If you receive a request from a data subject, it is advisable to consult the senior specialist for data protection by email at andmekaitse@ut.ee.

4. Publication of research results and data retention

Data protection is also important after the end of the research. In addition to ongoing compliance with data protection principles, it is important to know how to publish and store data. This chapter covers the main issues that may arise when sharing personal data and makes some recommendations.

Both the immediate outcome of the research and its dissemination and use by society are important for the public interest. Achieving the objectives of open science also requires that the research results and the data used to obtain them are permanently available to all interested parties for a long time. Article 10(1) of the EU Open Data Directive 2019/1024 states that "Member States shall support the availability of research data by adopting national policies and relevant actions aiming at making publicly funded research data openly available ('open access policies'), following the principle of 'open by default' and compatible with the FAIR principles."

Read more:

Chapter "Selection and long-term preservation" of the guidelines on creating the data management plan by the University of Tartu Library

4.1. How long can personal data used in research be stored?

There is some divergence between the principles of data protection and those of open science when it comes to storing personal data. The principles of minimisation and storage limitation imply that personal data should be processed for as short a period as possible and deleted or anonymised after the purposes for which they were collected have been fulfilled. However, it is in the interest of open science to ensure access to scientific data for as long as possible, at least as long as they are of value to researchers or society. Since it is also important for research to study historical events and trends, to compare past phenomena with those of today, or to understand processes more generally, there is no single time limit after which data lose their scientific value.

According to Article 5 (1) (e) of the GDPR, personal data may be kept in a personalised form beyond the purpose for which they were originally collected, provided that this is done solely for scientific research and that technical and organisational measures are taken to ensure the protection of individuals' privacy. The GDPR, therefore, offers more flexibility in storing personal data relating to research. At the same time, the European Data Protection Supervisor stresses in his preliminary opinion on data protection and scientific research (p. 18) that this special regime cannot be applied in such a way that the essence of the right to data protection is emptied out. For example, the privileges related to research are abused if personal data are retained indefinitely (derogation to the storage limitation principle), and at the same time, individuals' rights to their own information are limited (derogation to data subject rights). The research-related derogation can, therefore, only be invoked if the continued retention of personal data is legitimate, necessary and proportionate.

Therefore, the following conditions must be met to store personal data beyond the original purpose.

- **For research purposes only**: the derogation is granted for research purposes only and should not be used for the unlimited retention of personal data for other purposes which are of a private or commercial nature. A derogation is also made for archiving in the public interest and

for statistics, which means that personal data of archival value or relevant for statistical purposes may also be kept for a longer period. It is not always easy to draw a clear dividing line between research, archiving and statistics, so that both derogations may cover research.

- **Technical and organisational measures**: longer retention requires the secure storage of personal data. While during research, data availability is important, during storage, availability becomes less important and more secure storage solutions may be considered. For example, access to personal data by members of the research group may be restricted.
- **Data anonymisation should be considered**. Anonymised data could be stored indefinitely and also shared in an open data repository. However, anonymisation for storage purposes should be agreed upon at an early stage, and this intention should be clearly stated in the data management plan and in the information provided to research participants. If anonymisation is not possible, personal data should be pseudonymised, but in this case, the GDPR applies.
- **Only valuable data should be retained**. Data should be stored in accordance with the principle of minimisation, i.e. only the most relevant data should be retained, where it is necessary and justified to keep them in a personalised form for a long time. Since the research-related derogation always requires a balancing of different interests and needs, it may be helpful to delimit the data to be retained within the dataset, either to those of high long-term scientific value or those necessary for the validation of results. The GDPR does not allow storing data 'just in case'.
- **Storage must be transparent and fair.** Reliance on the derogation for the storage of personal data should be known already at the planning stage of the study. It would not be transparent and fair if only at the end of the study the research team decides to retain certain personal data for a longer period. The promises made to the data subject must also be respected: if the person is told that the data will be destroyed after the end of the study, it is not allowed to retain them. A difficult situation arises when, in the course of the research, it turns out that the data collected are much more valuable than expected, but the plan was to destroy all the data. Further use of personal data in new research and longer retention is possible under the GDPR, but the change to the original decision must be fair and transparent for the people.
- **Wherever possible, the storage of personal data should be considered as a separate purpose.** An appropriate legal basis for the new purpose must be found in this case.

4.2. In what form may personal data be disclosed?

Disclosure is defined as making personal data accessible to an unrestricted number of people, either on a public website, in a public database or elsewhere. Disclosure is only possible if the confidentiality of the data does not need to be guaranteed. As the security principle generally requires that the confidentiality of personal data must also be protected, derogation concerning the publication of personal data is possible in the case of research.

Disclosure does not concern situations where the person requests a copy of their data or where data are shared between research institutions.

4.2.1. Disclosure of personalised data

As disclosing personalised data is not usually necessary for research purposes, it should be considered as a separate purpose, which also needs an independent, unambiguous and clear legal basis. In the case of research, separate consent of the data subject is appropriate. If disclosure is not based on consent, the requirements of section 6 of the Personal Data Protection Act must be followed.

The disclosure of some personal data may also be necessary if the research has studied the creation of individuals, such as written or oral works (stories, biographies, media texts, etc.). In such cases, publication of the names of the authors of the texts and other works studied may be necessary and justified. The naming of authors and referring to them without their consent is permitted for the purposes of academic, artistic and literary expression (see section 5 of the Personal Data Protection Act).

4.2.2. Disclosure of pseudonymised data

As pseudonymised data are personal data, their disclosure must be based on a clear legal basis and be compatible with the purpose of the processing. Pseudonymisation only provides an additional safeguard against the identification of individuals but does not in itself give the right to disclose the data.

For example, quoting sentences from unpublished texts (e.g. transcripts of interviews) and referring to them with a pseudonym or a phrase that excludes personalisation (e.g. "...according to the doctoral student who participated in the interview...") can be considered as a disclosure of pseudonymised data. However, when quoting public texts (e.g. social media comments), it should be noted that the authors may be easily identifiable.

The researcher must keep the promises made to the research participants. If the researcher has promised anonymity, it is forbidden to publish pseudonymised data. However, if the researcher promises that the names of the subjects and the data that allow for their personalisation will be kept confidential, pseudonymised data may be disclosed if this is necessary for the purposes of the research. For example, it would not be purposeful for a researcher to publish pseudonymised interview transcripts that they find interesting or strange on a personal social media channel – such processing goes beyond the bounds of research and goes against the expectations of the data subject.

Pseudonymised data can be published in their entirety if there is a legal basis for doing so, data protection principles are respected and anonymisation of the dataset is not possible. However, in such cases, the university is responsible for the consequences of disclosure.

4.2.3. Disclosure of anonymised data

The preferred option is to disclose and share anonymised data, as these are no longer personal data and, therefore, no additional restrictions apply to their use. The GDPR and the Open Data Directive (EU) 2019/1024 favour the availability and wide use of scientific data, and the easiest way to ensure this is to anonymise research containing personal data.

4.3. With whom can personal data be shared when conducting research?

Personal data can be shared by providing a copy of the data or by granting access to the data. In either case, it is necessary to assess whether the controller has the right to share the personal data. The most typical cases of data sharing are set out below.

4.3.1. Data processing in a research group

Where the representatives of the university, i.e. the data controller, process personal data based on an employment relationship, there are no additional restrictions on such processing under the GDPR. However, general principles such as purpose limitation, minimisation and security (see 1.5) must be kept in mind, which do not allow personal data to be shared with every university employee. Thus, the need for each individual researcher to process personal data must be agreed upon within the research team or the university. For example, the data management plan can list the researchers who have access to personalised data, those who pseudonymise, those who keep the pseudonymisation secret and those who only process pseudonymised data. The data may then only be shared with the researchers identified in the data management plan.

Not all members of the research team may have an employment contract with the university. In this case, they must be authorised to process personal data.

4.3.2. Data processing in collaboration with several research institutions

For larger projects, the responsibility for personal data may be shared between research institutions, in which case attention needs to be paid to their roles and duties. As a general rule, agreements between research institutions should clearly specify the division of responsibilities for data protection and how personal data will be shared. Data protection principles must be respected, according to which personalised data are not accessible to all partners in the project but only to those whose task is to process them. Where possible, data should only be shared between partners in pseudonymised form and using secure solutions.

4.3.3. Data processing in cooperation between the supervisor and the supervisee

In the case of supervision within the university, there are no direct restrictions on sharing personal data, as both the supervisor and the supervisee are representatives of the controller, i.e. the university. Their cooperation and data sharing must be transparent: if the supervisor also sees personalised data, data subjects cannot be promised that no one but the supervisee will process personal data.

4.3.4. Sharing data with other researchers, publishers, repositories or the public

The preferred solution for sharing personal data with a wider audience is to anonymise them beforehand. If this is not possible for some reason, transferring personal data to third parties requires a legal basis – without this, personal data cannot be shared. In addition, it may be a good idea to ask the subjects for broad consent, i.e. for storing and sharing pseudonymised data for possible future research. If personal data are to be shared under section 6 of the Personal Data Protection Act without consent, this should be done in a pseudonymised form.

In addition, the recipient of the personal data must ensure adequate protection. To this end, a separate contract may be concluded with the recipient, setting out the conditions for using personal data, including the division of responsibility between the parties who will have access to the data. For example, it may be possible to place personal data in an open data repository for storage but restrict access to them, ensuring their confidentiality. Before using the repository, it should be ensured that it has data protection terms and conditions in place, which can be found in the terms of use of the service or in a separate document. Many scientific journals also ask authors to provide a data access statement (see 4.3.5).

While the responsibility for the unauthorised sharing of personal data lies with the researcher who has done so, this responsibility inevitably extends to the university as the data controller, which must record and report the data breach (see 3.5).

4.3.5. Conditions for sharing data with publishers

One prerequisite of open science is that the data used in a research project are accessible to all other researchers, either to validate previous research or for completely new studies. To this end, many publishers have established data sharing policies which, if accepted, require authors of articles to share data with other researchers.

It has become a new practice for publishers to require researchers to fill in a **data access statement** when publishing research articles, describing whether, where and how research data are available to other researchers. Three options for the use of personal data can be included:

- **If the data are already public**, the statement should indicate where they are located (in a repository of scientific data or another open data environment);
- **If the data are to be requested from the authors**, the statement should indicate that they will only be shared upon request. It is also possible to set additional conditions, for example, that the request can only be made by a researcher with a PhD or by the principal investigator of a study. Such conditions presuppose that there is a good reason why the data cannot simply be disclosed. The researcher requesting the data will be subject to a confidentiality agreement before the transfer of the data;
- **If the data cannot be shared**, the reasons why this is not possible or allowed must be explained in the statement. One reason may be the need to protect people's privacy.