Instructions for handling datasets containing sensitive personal data and key concepts

Reference: ATT aineistonhallinnan ohje sensitiivisille aineistoille -työryhmä (2018) Instructions for handling datasets containing personal data, Tuuliprojekti (document in Finnish)

Table of Contents

Instructions for handling datasets containing sensitive personal data

0. Introduction		Kommentit

Motivation	**THESE INSTRUCTIONS SUPPLEMENT THE NATIONAL DATA MANAGEMENT PLAN INSTRUCTIONS. READ THE INSTRUCTIONS SIDE BY SIDE!** These are instructions for drafting a **data management plan**, which is separate from a **research plan**. However, particularly in research which is based on collecting and analysing data, a research plan and data management plan may be closely interconnected and often overlap. The main difference between a research plan and a data management plan is that while the research plan describes which data will be used in the research, as well as why and how the data will be used, the data management plan lays out how the data will be managed, and how further use of the data is enabled in the course of research. These instructions supplement the general data management plan guidelines as they pertain to datasets which contain **sensitive personal data.** All of the protective measures described in these instructions will not be relevant if the personal data is not deemed sensitive. Personal data means any information relating to an identified or identifiable natural person. An identifiable natural person is one who can be identified, directly or indirectly, in particular by reference to an identifier such as a name, an identification number, location data, an online identifier, or to one or more factors specific to the physical, physiological, genetic, mental, economic, cultural or social identity of that natural person. Special categories of personal data which is particularly sensitive (Articles 9 and 10 of the GDPR) include: 1. Racial or ethnic origin 2. Political opinion 3. Religion or beliefs 4. Trade union membership 5. Genetic or biometric data processed for the purpose of uniquely identifying a person 6. Health information 7. Sexual behaviour or orientation 8. Criminal convictions and offences **Purely to make this guide easier to understand, we call "sensitive personal data" the data described above. However the exact legal term is "special categories of personal data".** **The processing of personal data is regulated by legislation.** The legislation governing the processing of personal data is the **EU's General Data Protection Regulation (GDPR)**, along with the **Data Protection Act** that supplements it. The purpose of the new legislation is to improve people's opportunities to decide how information about them is processed, and it also has implications for how personal data is processed in research. New features include the accountability requirement, which means the controller or processor of the personal data must in the future demonstrate in writing that they comply with data protection legislation and the principles of processing personal data while ensuring the legal rights of the data subjects. In addition, there are changes to the rules governing how personal data collected with the consent of the subject can be used. There are also **organisation-specific instructions** for many stages of the processing of personal data which must be followed. Data management planning is particularly important when processing datasets containing personal data, as it allows you to protect your rights and the rights of your organisation, as well as the rights of your research subjects. The breach of data protection legislation may result in **administrative sanctions, criminal liability and liability for damages.** Letting personal data fall into the wrong hands may cause **serious damage to the research subject.** **Further information:** The Data Protection Ombudsman's office is currently drafting instructions on applying the new data protection legislation: https://tietosuoja.fi/en/home	Kommentoitava versio löytyy täältä: https://wiki.helsinki.fi/x/EvOKDw
1. General description of data		

1.1 What kinds of data is your research based on? What data will be collected, produced or reused? What file formats will the data be in?	The **data management plan** should describe the kind of personal data the collection and analysis methods generate. The justifications for the research and the reasons for collecting and processing personal data should be included in the **research plan.** Describe all relevant data sources in the data management plan. For example, list the people or groups of people, authorities and registers involved in the research. For each data source: • Itemise all data that includes *personal data* or *sensitive personal data*. • If the data source is a register or a statistic, also indicate its *controller*. • Indicate who, or which organisation, is the data file *controller* of the data you collect or produce. Please note that **when you collect personal data or sensitive information, you must also ensure the security of the media used to collect and transport the data**. A more detailed description of this is included in section 4.1.	
1.2 How will the consistency and quality of data be controlled?	Consider the quality of the data throughout its life cycle, from collection to publication and archiving. What are the biggest risks and how will they be managed? Does the collection of data which contains personal information feature elements that require special attention in relation to the quality of the data? (Information security will be covered in section 4.1) **Tips** • Consider when the data should be protected with a code or whether it should be anonymised. • Remember the difference between anonymised and pseudonymised data. • Consider whether *anonymisation* or *pseudonymisation* will impact the quality of the data. Will the data still be useful after anonymisation? • Remember to ensure that no valuable information is lost if the data is made less specific. • Recording metadata and using metadata standards are also quality measures and should be entered in more detail in section 3, "Documentation and metadata" of your plan.	
2. Ethical and Legal Compliance		

2.1 What ethical issues are related to your data management, for example, in handling sensitive data, protecting the identity of participants, or gaining consent for data sharing?	Indicate in your plan who, or what organisation, is the data file *controller* of the data you collect or produce. Also indicate who the **processors** are who process the personal data on behalf of the controller. **The processing of personal data** means any operation which is performed on personal data, such as collection, recording, organisation, use, storage, adaptation or alteration, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction. Data processing also includes cases in which parties outside the organisation or research project analyse samples. Processing agreements must be drafted with such third parties. The processor must take protective steps to safeguard the rights of the data subject. Such protective measures include: • *pseudonymisation* • *anonymisation* • sufficient *safeguards*: technical restrictions, use monitoring, **described in section 4 of the plan** • training, instructions, regulations, commitments and agreements • processes, practices and certificates • data encryption • audits **Data protection impact assessment** Your plan should indicate how the impact assessment will be carried out. The purpose of the impact assessment is to describe how the personal data will be processed. Assess the necessity and proportionality of the processing and assess the risks resulting from the processing as well as measures necessary to address the risks. Impact assessment is required when the processing of personal data is likely to carry a high risk. The purpose of the impact assessment is to help the controller comply with the requirements of the GDPR and to demonstrate this compliance. Data protection impact assessment should begin as early as possible when the processing of personal data is being planned. The assessment must be constantly monitored and updated whenever necessary. **Tips** • Refer to the data protection instructions for your organisation. • Refer to your organisation's instructions on processing contracts. • Refer to the impact assessment instructions of your organisation and the office of the Data Protection Ombudsman. **Links:** • Data protection documentation: http://www.tietosuoja.fi/en/index/materiaalia.html • The EU data protection reform: https://tietosuoja.fi/en/organisations • Anonymisation and identifiers: http://www.fsd.uta.fi/aineistonhallinta/en/anonymisation-and-identifiers.html • Finnish Social Science Data Archive Data Management Guidelines: Informing Research Participants: http://www.fsd.uta.fi/aineistonhallinta/en/informing-research-participants.html • Data Protection Ombudsman's office instructions for impact assessment (in Finnish): http://www.tietosuoja.fi/fi/index/euntietosuojauudistus/ohjeitarekisterinpitajalle/vaikutustenarviointi.html	
2.2 How will data ownership, copyright and Intellectual Property Right (IPR) issues be managed? Are there any copyrights, licenses or other restrictions which prevent you from using or sharing the data?	The ownership, copyright and intellectual property rights of the data must also be recognised. This is particularly important for sensitive data of any kind. **Tips** • Carefully read the terms of use for all of the IT services you use. • Written agreements regarding data ownership, use rights and publication authorship help ensure data protection.	
3. Documentation and metadata		

3.1 How will you document your data in order to make it findable, accessible, interoperable and re-usable for you and others? What kind of metadata standards, README files or other documentation will you use to help others to understand and use your data?	**Tips** • In the description of variables, mention whether the variable contains personal or sensitive data. Refer to, e.g., the Data Management Guidelines. • Even if your research data contains personal data, you may publish the metadata if it contains no identifiers which could be used to identify the research subject.	
4. Storage and backup during the research project		
4.1 Where will your data be stored, and how will it be backed up?	If your research involves collecting or using personal data or sensitive personal data: • Consider the requirements of the party disclosing or transmitting the data as early as possible • Draft the statutory risk assessment, indicating the information security measures required Data protection measures include: • Backup copies: ensure the ability to recover after a systems failure • Access control: who is granted access and on what grounds, how is the access restricted, **this is described in more detail in section 4.2** • Encryption: whenever necessary. Encryption is especially recommended for mobile devices, laptop computers and external storage devices. • Monitoring: both a technical log and monitoring of data processing and use, **described in more detail in section 4.2.** • Protecting the technical environment: how can the processing environment be protected from third parties • Personnel security: orientation of research group members, data protection and information security training, instructions and shared practices • Facility security: locks on work spaces, storage furniture, camera surveillance and access control, **described in more detail in section 4.2.** **Tips** • Whenever possible, use the protected processing environments recommended by the controllers. • Remember that the transfer of personal data outside the EU and EEA has been restricted. • Bear in mind that consent forms also contain personal data. ***Links:*** • *The Finnish Social Science Data Archive Data Management Guidelines:* http://www.fsd.uta.fi/aineistonhallinta/en/physical-data-storage.html • *Health and Welfare Data Portal:* http://hytedata.fi/en/ • *Finnish Communications Regulatory Authority / National Cyber Security Centre: Cyber threats in the health care industry (document in Finnish): https://www.viestintavirasto.fi/attachments/tietoturva/Terveydenhuoltoalan_kyberuhkia.pdf* • *Cloud Guide for institutions of higher education:* https://wiki.eduuni.fi/display/cloudguide/Cloud+Guide	

4.2 Who will be responsible for controlling access to your data, and how will secured access be controlled?	**Access control:** who is granted access and on what grounds, how is the access restricted, and who is responsible for access control?
	
	• A person must be designated to be in charge of access control
	• A list of granted access rights and users must be drafted
	• Access is only granted when needed, and the access must be as limited as possible.
	• The user's need and basis for accessing the data must be inspected before granting access
	• A system must be in place for revoking and deleting access rights
	
	Monitoring: this means both a technical log and procedures for monitoring the processing and use of the data.
	
	• Consider how the use of the data will be monitored over the course of the research.
	  • Where and in what ways will the data be processed?
	  • Where and for whom can it be copied?
	  • Who can transfer data outside the research group and on what grounds? Remember that this must be in line with the consent from the data subjects if the data has been made available based on consent.
	  • Examine whether, and describe how, the technical tools used can keep a log of who used which data and when. Ask your organisation's IT support for use and change logging.
	
	Facility security: locks on work spaces, storage furniture, camera surveillance and access control.
	
	• A person must be designated to be in charge of access control
	• A list of holders of access rights and keys must be drafted
	• Which doors are locked or are lockable between the work space and the outside?
	• Are there theft-proof storage facilities or furniture available in the work spaces for documents, other analogue material and external storage devices?
	• Is camera surveillance available?
	
	Tips
	You will need:
	
	• A document that complies with the accountability requirement of the GDPR
	• A statement of data protection measures
	
	Links:
	
	• *Finnish Communications Regulatory Authority / National Cyber Security Centre: Instructions on recording and using log data (document in Finnish):https://www.viestintavirasto.fi/attachments/tietoturva/Lokitusohje.pdf*	

5. Opening, publishing and archiving the data after the research project		

5.1 What part of the data can be made openly available or published? Where and when will the data, or its metadata, be made available?	Material containing personal data can only be released once it has been anonymised. Pseudonymised data still constitutes personal data and can consequently not be released. Material which contains personal data may, however, be shared with interested parties upon request for the purpose cited in the original basis for processing. The basis for processing material containing personal data, for example a statutory reason or consent, may restrict the ways the data can be used later. Acceptable ways to release or publish material which contains personal data include: 1. The data is anonymised and released into a data archive with an appropriate level of data protection 2. Only the metadata for the material is published in a suitable research database or data repository. **Tips** • Key metadata for material containing personal data should be released even if the material itself cannot be. • Pseudonymised data is still personal data, and cannot be released for further use. However, further use of the material may be possible by request. • Further use of the material may require that new consent be requested from the research subject. **Links:** • Anonymisation and identifiers: http://www.fsd.uta.fi/aineistonhallinta/en/anonymisation-and-identifiers.html	
5.2 Where will data with longterm value be archived, and for how long?	When drafting an archiving plan, it is important to consider which parts of the material will be *archived*, and for what period of time. It is also important to decide which parts will be destroyed and how this can be done securely. Traditionally, the recommendation has been to destroy all sensitive data after the research project, as storing it carries risks and requires special arrangements. Other unnecessary files and intermediate files generated by IT systems must also be deleted once they are no longer necessary. Just deleting a file and emptying the recycle bin on the computer does not mean that the file has been permanently destroyed. It is possible to retrieve deleted files even after the hard disk has been reformatted. A variety of applications exist for permanently destroying data, based on overwriting data or magnetising the hard disk. It is also possible to mechanically crush the storage device so that it cannot be read. Archiving material that contains sensitive personal data requires permission from the National Archives, and the data must be *minimised* before archiving. Any later use of such material requires a research permit. **Tips** • Please remember that the anonymisation and destruction or archiving of the data must be done by the deadline of the research permit. • Genuine anonymisation requires that there is no possibility of either direct or indirect identification, and that the code key is destroyed. • Data relating to samples may be archived in a biobank. • Many universities and public authorities have their own internal guidelines for destroying storage devices. **Links:** • FSD: Data disposal: http://www.fsd.uta.fi/aineistonhallinta/en/physical-data-storage.html#disposal	
5.3 Estimate the time and effort required for preparing the data in order to publish or to archive it.	When evaluating the costs associated with the management of sensitive data, consider: • the costs of anonymising data (the time and programs required) • the technical requirements of a higher level of security	

Key concepts

Anonymisation
Finnish Social Science Data Archive Data Management Guidelines, http://www.fsd.uta.fi/aineistonhallinta/en/anonymisation-and-identifiers.html: Data are anonymised if characteristic factors (for instance, indirect identifiers when linked together) are the same for several individuals and if any particular individual cannot be identified with reasonable effort. The assessment of how identifiable the data of a dataset are and how they can be anonymised is always done on a case-by-case basis.
Sensitive personal data
Purely to make this guide easier to understand, we call "sensitive personal data" the data described below. However the exact legal term is "special categories of personal data".
Special categories of personal data (Articles 9 and 10 of the GDPR) include:

- Racial or ethnic origin
- Political opinion
- Religion or beliefs
- Trade union membership
- Genetic or biometric data processed for the purpose of uniquely identifying a person
- Health information
- Sexual behaviour or orientation
- Criminal convictions and offences

Archiving
Archiving is a means to ensure that documents are recorded and that they remain usable, and also a means to arrange the information service associated with the documents (Archives Act).
Data archive
A data archive is used to store research data for use during the project and for long-term storage.
Data repository
This term is used as an umbrella concept for various levels of databases into which data can be stored and described. The difference between a data repository and a data archive is that the latter is considered to be a database for long-term data storage. Conversely, a repository carries no implications of long-term preservation. Some repositories only contain metadata and not the data itself. Data repositories are listed in the re3data service.
Ethical review
A statement issued by the research ethics committee regarding whether the research complies with general ethical rules.
Personal data file
A personal data file means a set of personal data, connected by a common use and processed fully or partially automatically or sorted into a card index, directory or other manually accessible form so that the data pertaining to a given person can be retrieved easily and at reasonable cost (Personal Data Act).
Personal data
All data related to an identified or identifiable person are personal data.
In other words, data that can be used to identify a person directly or indirectly, such as by combining an individual data item with some other piece of data that enables identification, are personal data. Persons can be identified by their name, personal identity code or some other specific factor.
Source: https://tietosuoja.fi/en/what-is-personal-data
Processor
The processor of the personal data processes the data for the controller. The processor must take protective steps to safeguard the rights of the data subject.
Processing of personal data
This term means any operation which is performed on personal data, such as collection, recording, organisation, use, storage, adaptation or alteration, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction.
Lawfulness of processing
A legal reason must always be demonstrated for the processing of personal data. This reason must be defined before the processing begins. Once the processing of personal data has been linked to a specific reason, this reason can no longer be changed to another one.
The GDPR lists six reasons which enable the processing of personal data:

- consent from the data subject
- contract
- the controller must comply with a legal obligation
- the protection of vital interests
- public interest or official authority
- legitimate interests of the controller or a third party.

Source: https://tietosuoja.fi/en/when-is-the-processing-of-personal-data-permitted

Processing procedures

The collection, recording, organisation, structuring, storage, adaptation or alteration, retrieval, consultation, use, disclosure by transmission, dissemination or otherwise making available, alignment or combination, restriction, erasure or destruction as well as other potential forms of processing.

Purpose / purpose of the processing

On a general level, the purpose is academic research. A more detailed purpose is described in the data management plan and in the research plan.

Metadata

Data about data, i.e., descriptive and defining data about a data resource or content unit

Minimisation

The level of detail in the personal data must fit the purposes of the processing

Pseudonymisation

Pseudonymisation means the processing of personal data in such a manner that the personal data can no longer be attributed to a specific person without the use of additional information. Such additional information must be kept carefully separate from personal data.

Source: https://tietosuoja.fi/en/pseudonymised-and-anonymised-data

Controller The controller is a person, corporation, institution or foundation, or a number of these, for whose use a personal data file is set up and who is entitled to determine the use of the file, or who has been designated as a controller by legislation.

*"*The controller shall implement appropriate technical and organisational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed. That obligation applies to the amount of personal data collected, the extent of their processing, the period of their storage and their accessibility. In particular, such measures shall ensure that by default personal data are not made accessible without the individual's intervention to an indefinite number of natural persons.*" (Article 25, EU GDPR, "Data protection by design and by default".*

Safeguards

Measures to be taken in addition to those required by data protection legislation, including national special legislation, the appointment of a data protection officer, impact assessment, audits, collecting log information, etc.

Consent

Consent means any voluntary, detailed and conscious expression of will, whereby the data subject approves the processing of personal data. (Source: http://www.tietosuoja.fi/fi/index/sanasto.html)

Transparency

In this context, transparency means openness towards the research subjects, who must be informed whenever possible of the research and the ways the data will be used.

Records of processing activities

The controller and the person processing the personal data on behalf of the controller must maintain records of the processing activities under their responsibility.

Period for which the personal data is stored

Includes the planned deletion dates of different data groups or the criteria to be used for determining the storage periods. The periods of storage are related to the principles of data minimisation and storage limitation. The determined period for which the personal data is stored must indicate how long the data of the data subject will be processed. It is not sufficient to state that the personal data will be stored for as long as necessary to reach certain legal objectives.

Data containing identifiers

Finnish Social Science Data Archive Data Management Guidelines, http://www.fsd.uta.fi/aineistonhallinta/en/anonymisation-and-identifiers.html: Data is considered to contain identifiers if it can be used to identify an individual person. This identification can be made on the basis of factors specific to the physical, psychological, mental, economic, cultural or social identity of an individual or individuals.

Impact assessment

The assessment of the impact of the processing on the protection of the personal data.