WeNet Guiding Material
Data Minimization

How to cite this resource:

Schelenz, Laura. (2020). WeNet Guiding Material on Data Minimization, EU-funded project "WeNet – The Internet of Us", available online.

Part 1:

Data Minimization in Practice – Advice and Further Resources

Why Data Minimization?

The **EU General Data Protection Regulation** (GDPR) states that "personal data shall be … adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed ('data minimisation')" (Article 5.1.c).

When it comes to the **processing of personal data** of natural persons, the data should be limited to what is necessary in order to fulfill the purpose of the processing!

This means that personal data should only be collected and processed if it **immediately fulfills the purpose of the processing**. Even if more personal data was available, it would be disregarded in the data processing if the processing of less information achieves the intended goal.

Personal data is any information that identifies a natural person or, as they are usually referred to, data subjects. Examples of personal data include "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" GDPR Article 4.1

Sensitive data (also called "special categories of data") is such personal data that relates to the fundamental rights of a person and therefore requires higher levels of protection. Examples include data revealing "racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, genetic data, biometric data, data concerning health or data concerning sex life or sexual orientation" GDPR Article 9.1

The goal of data minimization is to **reduce the amount of data relating to a person (and especially such information that is sensitive) to a minimum** in order to protect that person (or 'data subject') from violations of their fundamental rights.

Preliminary Method for Reducing Data Types BEFORE Data Collection

From your list of possible data types that can potentially be collected, we suggest that – as an exercise – you pick a maximum of 10 items that are essential to creating your system.

The process to pick the ten items may be executed in three steps:

1. All stakeholders in the design process of your system specify their respective needs for types of data. They filter data types needed for their specific tasks from the long list of potentially collectable data (initial application of data minimization principle).

Involving all stakeholders (e.g. all project partners within a project, all interested parties in the dataset) ensures that the "minimized data" covers the different areas in a project, and thus fulfills the overall purpose of the project.

2. From the reduced list of data types, each stakeholder picks only 1 or 2 or x amount of data types that are absolutely necessary for the design and development of the system (strict application of the data minimization principle).

   Example: If there are 5 stakeholders in a project, and the goal is to arrive at a list of 10 items, every stakeholder picks 2 types of data that are essential to their activities.

   This is done by intensive discussion with the team, using the following matrix as a support tool. The follow rule of thumb may be applied: If the data type is rather sensitive and less important to fulfill the purpose of the project, the data type should be excluded from the final list of collectable data (see field 2 in the table below). If the data type is less sensitive and rather important to fulfill the purpose of the project, the data type should be included in the final list and does not need further discussion (see field 3 in the table below).

   Particular attention must be paid to fields 1 and 4. Data types that fall into these categories (rather sensitive + rather important and less sensitive + less important) should be discussed among the stakeholders. What are the ethical implications of using these data types? Are the risks to the data subjects worth the benefit for the project? Please reflect and come up with a qualitative judgment through joint evaluation of the data types in fields 1 and 4.

Data	Rather important	Less important
Rather sensitive	1	2
Less sensitive	3	4

3. In the final step, the stakeholders fill in the template below, which represents the consensus of the community regarding the collection of data. Fill in the table by specifying a) why the respective data type is essential for creating your systems, b) the diversity dimension that is reflected by the data type (for diversity-aware technology), c) whether the data type is rather sensitive or less sensitive, and d) what potential ethical considerations have to be taken into account when collecting this data type.

WeNet Guiding Material
Data Minimization

List of data types for design and development of a system (data minimization applied)

	data type	why important? (a)	diversity dimension (b)	level of sensitivity (c)	potential ethical concerns (d)
1					
2					
3					
4					
5					
6					
7					
8					
9					
10					

Part 2:

Data Minimization Case Study: Sensor Data for Diversity-Aware Technology

Task: we want to develop a platform that leverages the diversity of technology users to their advantage. We therefore need to collect data that represents the diversity of users and identifies their needs and preferences. We have the opportunity to collect sensor data from users' smartphones. ***What types of data should we collect while at the same time adhering to important legal and ethical principles?***

This case study builds on a data minimization exercise conducted in the WeNet project.

The methodology used to come up with a reduced list of sensor data types included three steps: 1) initial filtering of data types from a long list of sensor data types (each project partner), 2) strict application of data minimization by asking each project partner to pick 2 types of sensor data, 3) discussion of the remaining sensor data types and consensus regarding a final list of sensor data items.

The sensor data identified as ***essential to developing diversity-aware technology for social interaction*** includes the following:

- **Location**
- **Physical activity (via accelerometer)**
- **History of interactions (via calls, texts, social media)**
- **Touch event**
- **Interruptions (via flight mode and any indication that notifications are set on or off)**
- **Contact list**

Particularly ***sensitive*** are the data types location, history of interaction, and contact list, because they potentially reveal information about the user's personal identity, e.g. ethnicity, health status, sexual orientation, political and religious affiliations. When using these sensor data types, researchers and developers should ask whether they can fulfill the purpose of their task by collecting less sensitive data or limit the scope of their collection further.

Location allows data scientists and researchers to tailor content/events/promotions to the particular geographical region of the user. It also reveals routines of data subjects, e.g. the daily commute to and from work or school, visits to friends, family, doctors, shops, etc. Diversity-aware technology may leverage location to determine variations of routines among individuals and offer them suitable recommendations based on their routine behavior. Location information is sensitive, as it reveals not only daily routines, but may be used as proxy data for socio-economic status, health, and race/ethnicity in countries that have highly segregated urban and rural structures.

Physical activity is interesting to developers who build health-related services. For a diversity-aware social network of engaged users, acceleration tells us about the movement of a user, and allows us to conclude the level of activity, fitness or strength. It is thus possible to cater particular content to users who are less or more interested in sports – and who are in general more active or passive. Acceleration is not sensitive as the information does not relate to fundamental rights. Yet, acceleration may reveal facts about sleeping patterns and transportation mode.

The ***history of interactions***, e.g. incoming and outgoing calls and texts (without audio), are crucial to reconstruct a user network and recompile this social network in the service offered. It may also be possible to collect information from social media sites used by the data subject, e.g. Facebook posts and friend lists. There seems to be endless supply of information on a person's social network if data includes phone histories and social media. It is then possible to learn about a person's faith practices (the community they engage with), a person's health (the health professionals they are in contact with), a person's sexual orientation (if openly seeking sex partners or being politically active in the LGBTIQ community), a person's political opinion (by consuming and sharing news, communication with local politicians, parties), and trade union membership (e.g. via group membership on Facebook). This data is *highly sensitive*. It may be warranted to further examine the "history interaction" and narrowly determine the scope of data that is required to fulfill the purpose of the data collection. Do we really have to parse Facebook likes and shares? Or would it suffice to use the "friend list" in Facebook to determine initial relationships/networks?

Touch event refers to the frequency and type of touch on the phone. In theory, the concrete movements of a finger can be reconstructed by combining analyses of different touches. However, the goal is to understand first and foremost how engaged a user is with their smartphone. This data is not sensitive.

Interruptions, such as any modalities that prevent the user from receiving messages: If we collect flight mode as a data type, we understand the reasons for a user not to engage with a certain message – they might not have received it. Interruptions reveal potential sources of disengagement. This information is not sensitive.

WeNet Guiding Material
Data Minimization

The **contact list** in a phone may be used to draw conclusions about the social network of the person. It allows designers to recompile the list of contacts in service offered and make suggestions for interaction based on previous acquaintance. This information can be rather sensitive because it may reveal contact with religious, political, or trade union organizations, as well as health status if users have the information of specific doctors in their contact list.

Particular concern should be raised about the possibility to combine these selected data types: Even if we use data minimization and limit the sensor data to a few essential types, we still may be able to “learn everything” about a person by combining different data points and running combined analyses. A data type may then not be sensitive by itself, but combined with other data types, we can infer sensitive information about the person.

The collection of sensor data for diversity-aware technology requires the ***informed consent*** of the data subject. This also means explaining to the user the potential privacy risks that arise from singular or combined use of sensor data types. Moreover, ***anonymization*** is key to protecting the privacy of a data subject, and even anonymized datasets should be shared only with immediately involved partners of a technology development project.

A tension remains between the quality of service and data protection. If users refuse to provide information, the quality of the services catered to them may be compromised. Users have to make an informed decision about the value they attribute to the services and to their privacy. This decision can be made autonomously only if the system is transparent about the kind of data collected and its usage. Transparency about data collection and protection measures (such as data minimization) also increase trust in the designers and the system.