Handouts on the processing of personal data for the purposes of language research and archiving of language resources under the General Data Protection Regulation

Version 1.0, September 2021

Author: Dr. iur. Paweł Kamocki (IDS Mannheim)

These guidelines were produced within the CLARIAH-DE project funded by the BMBF

SPONSORED BY THE

Federal Ministry of Education and Research

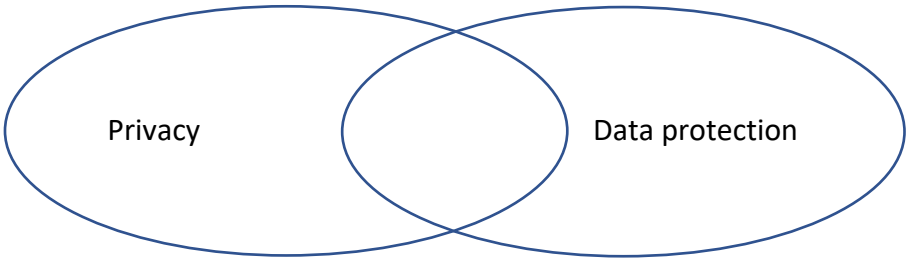
FUNDING REFERENCE NUMBER
01UG1910 A to I

This work is licensed under a Creative Commons Attribution 4.0 International License.

DOI: *https://doi.org/10.14618/ids-pub-10695*

Table of Contents

Introduction: Privacy vs. Data Protection



Privacy in its many aspects is protected by various legal texts (e.g. the Basic Law, Civil Code, Criminal Code, or even the Law on Copyright in artistic and photographic works (*KunstUrhG*), which protects image rights). **Data protection** law, which governs the processing of information about individuals (personal data), also serves to protect their privacy. However, some information referring to the public sphere of an individual's life (e.g. the fact that X is a mayor of Smallville) may still be considered **personal data** (see below), and as such fall within the scope of data protection rules. In this sense, data protection laws concern information that is not private.

Therefore, privacy and data protection, although closely related, are **distinct notions**: one can violate someone else's privacy without processing his or her personal data (e.g. simply by knocking at one's door at night, uninvited), and *vice versa*: one can violate data protection rules without violating privacy.

The following handouts focus exclusively on data protection rules, and specifically on **the General Data Protection Regulation (GDPR).** However, please keep in mind that compliance with the GDPR is not the only aspect of protecting privacy of individuals in research projects. Other rules, such as academic ethics and community standards (such as **CARE**) also need to be observed.

To get a full picture of data protection rules, it is advised to read all the handouts, even if at first they seem irrelevant for your project. When you are familiar with the general framework, you can re-read the handouts that are most relevant for you, and consider reading some of the documents that they refer to.

Sources of data protection law

Data protection law is defined in a number of texts (sources of law), which form a hierarchy.

European texts

Council of Europe's Convention for the Protection of Individuals with Regard to the Processing of Personal Data (**Convention 108**) – open to signature in **1981**, 55 ratifiers (all 47 members of the Council of Europe + e.g. Mexico, Senegal, Uruguay…)

EU Data Protection Directive 1995 – inspired by the Convention 108, replaced by the GDPR in 2018

Article 8 of the 2000 **Charter of Fundamental Rights of the European Union** – raises data protection to the rank of a fundamental right (distinct from the right to privacy) in the EU

EU e-Privacy Directive 2002 – contains specific rules on electronic communications (e.g. cookies, traffic data, unsolicited communications), expected to be replaced in near future by a regulation

EU General Data Protection Regulation (GDPR), adopted in 2016 and entered into application on May 25, 2018, it replaced the 1995 Personal Data Directive. As a regulation (and unlike a directive), the GDPR applies directly in all EU Member States, and supersedes national rules. The GDPR also applies in other countries of the European Economic Area (Iceland, Lichtenstein and Norway).

National texts

The GDPR supersedes national laws on data protection in the EU; however, the GDPR expressly leaves some issues (including some of those related to research and archiving) to the national legislators. For this reason, national texts on data protection remain an important source of data protection rules.

In Germany, the **BDSG** (*Bundesdatenschutzgesetz*, Federal Data Protection Act) applies to the processing of personal data by **private bodies** and by **public bodies of the Federation**. Each federal state (*Land*) also has its own **LDSG** *Landesdatenschutzgesetz*, which applies to the processing of personal data by public bodies of the *Land* (such as **universities**).

As these texts may differ, especially in the area of research and archiving, it is important to know which text applies to your institution (BDSG vs. LDSG). The following handouts will focus on the GDPR, i.e. the rules that apply uniformly through the EU. For certain specific rules, examples from the BDSG and LDSG of Baden-Württemberg will be used.

NOTE: The GDPR entered into application on May 25, 2018, and it does not apply retroactively, i.e. to processing of personal data that took place before that date. However, if you (continue to) process (even just store!) personal data collected before May 25, 2018 after that date, you have to comply with the GDPR. In this sense, the GDPR also applies to **legacy data**.

Basic Concepts 1: personal data, sensitive data, processing…

Personal Data is defined very broadly as *"any information related to an identified or identifiable natural person"* (Article 4 of the GDPR). This definition can be analyzed into four elements:

1. **Any information** – regardless of its form (analogue/digital, video/audio/text, etc.) or content (facts/opinions, true/false);
2. **Related to** (a person) – information relates to a person when:
   - It is about a person (the **content** element) OR
   - It can be used to evaluate or influence a person (the **purpose** element) OR
   - It has an impact on the person's rights and interests (the **result** element).

   NOTE: The Court of Justice of the EU ruled that exam scripts (answers to exam questions and evaluations thereof) constitute personal data of the candidate (case C-434/16, Nowak).
3. **An identified or identifiable** (person)
   - Identified – singled out from a group
   - Identifiable – possible to identify:
     i. Directly (via a username, ID number) OR
     ii. Indirectly – by a unique combination of elements: e.g. name + home town + date of birth, or geographical area + sex + languages spoken (e.g. for immigrants speaking endangered languages)
   - Information constitutes personal data if the person it relates to can be identified **by means reasonably likely to be used** (e.g., today's cutting edge face recognition technology may not be likely to be used to identify people on an aerial photograph; however, this may change overtime).

   NOTE 1: With social media, it is increasingly easy to uniquely identify a person based on relatively little information.

   NOTE 2: Information should be regarded as personal data even if the controller himself does not have access to all the elements necessary to identify the person. The Court of Justice of the EU ruled that **a dynamic IP address is personal data** because the **police** can obtain from the service provider the information necessary to identify the person behind the address and is **reasonably likely** to do this (case C-582/14, Breyer).

   NOTE 3: An American scholar, Latanya Sweeney, demonstrated that over 87% of US population can be uniquely identified by combining three seemingly non-identifying elements: gender, date of birth and ZIP (postal) code.
4. **Natural person**, i.e. a living individual. Information about the deceased is no longer "their" personal data, but it can still relate to other living individuals (e.g. the fact that one's grandfather died of cancer in a young age can have an impact on his or her situation, e.g. the cost of his life insurance).
   The person that the data refer to is called a **data subject.**

EXAMPLES of personal data:
- Interviews (potentially personal data of both the interview and the interviewee);
- Student essays;
- Voice recordings (both voice and the content of what is being said can be used to identify the speaker);
- Photographs and video recordings;
- Usernames, passwords, IP addresses;
- E-mail addresses, phone numbers (professional and personal);
- Phone records, sent e-mails
- Student essays…

And many, many more.

For a more in-depth analysis of the concept of personal data, see the WP29 Guidelines.

Special categories of personal data (a.k.a. sensitive data) are subject to stricter rules (Article 9 of the GDPR). These are data revealing:

- Racial or ethnic origin;
- Political opinions;
- Religious or philosophical beliefs;
- Trade union membership,

as well as:

- Genetic data;
- Health data;
- Data about sex life and sexual orientation;
- Biometric data, but only if they are processed for the purposes of uniquely identifying a person

NOTE: Voice, body language or facial expressions are biometric data, as they can be used to uniquely identify a person. However, if such data are not used for this purpose, they should not be regarded as sensitive data.

Processing of personal data is "any operation or set of operations which is performed on personal data or on sets of personal data, whether or not by automated means" (Article 4 of the GDPR). Processing includes among others collection, recording, storage, annotation, consultation, transmission, combination, anonymization, erasure or destruction.

Basic Concepts 2: anonymization and pseudonymization

Personal data can be anonymized or pseudonymized.

Anonymization consists of transforming personal data in such a way that they no longer refer to a person that is identifiable by means reasonably likely to be used.

Data protection rules do not apply to anonymized data. However, the standard for anonymization is high.

Anonymization has to be **permanent** and **irreversible**, as opposed to pseudonymization.

NOTE: With technological development, data that are anonymized now can be used to identify people in the future. Therefore, the results of anonymization should be periodically reviewed.

The appropriate technique for data anonymization depends on the nature of the data and the context. There is no one bulletproof solution, and the larger and more complex the dataset, the more sophisticated anonymization techniques would have to be used.

Some known anonymization techniques include randomization, generalization (t-closeness, k-anonymity, l-diversity) and noise addition. For an overview of these techniques and their assessment, see WP29 Guidelines on the subject.

Anonymization of language data can significantly reduce their research value.

In general, simply deleting named entities from text data, or simply 'beeping' them in speech data is rarely enough to achieve anonymization. Although data pre-processed in such a way will often not be anonymous, the pre-processing can potentially be regarded as a **safeguard** for rights and freedoms of data subjects.

Pseudonymization, on the other hand, consists of transforming personal data in such a way that they can no longer be attributed to the data subject **without additional information,** which is kept separately in a secure environment.

EXAMPLE: In a research dataset the participants are labelled with pseudonyms, i.e. false names or ID numbers (participant 1, etc.), whereas all potentially identifying information about the participants (their name, age, gender, city of origin…) is stored separately on PI's computer; the PI also has "the key", i.e. the information allowing to attribute 'real' personal data to pseudonyms.

Pseudonymized data are still to be regarded as personal data (because the 'key' exists and is reasonably likely to be used to identify participants), and their processing should follow the GDPR. However, pseudonymization is a good **safeguard** for the rights and freedoms of data subjects, especially in research context.

Basic Concepts 3: Who is Who? Data subject, controller, processor…

Data Subject – the natural person (living individual) that the data refer to. The data subject has certain rights with regards to his or her data.

Data Controller – the person or organization who (alone or together with others) decides about the means and purposes of data processing, i.e. *how* and *why* the data are processed. The data controller is accountable for compliance with the GDPR.

NOTE: In research context, generally the **institution** (e.g. university) is considered the controller of personal data processed by researchers, as they provide the infrastructure (hardware, software, physical and digital archives) and define, via their internal policies, how personal data should be processed. However, there are good arguments to consider independent researchers (e.g. the PI) as **joint controllers** alongside the institution. For advice on identifying the Data Controller in your project, contact your hierarchy and/or the Data Protection Officer at your institution.

Joint Controllers – several controllers who jointly decide about the means and purposes of data processing. In research projects carried out jointly be several institutions, all these institutions are likely to be joint controllers.

European Data Protection Board's Guidelines 07/2020 on the concepts of controller and processor v.2 (adopted on 7 July 2021) contain the following EXAMPLE (p. 22):

Several research institutes decide to participate in a specific joint research project and to use to that end the existing platform of one of the institutes involved in the project. Each institute feeds personal data it already holds into the platform for the purpose of the joint research and uses the data provided by others through the platform for carrying out the research. In this case, ***all institutes qualify as joint controllers*** *for the personal data processing that is done by storing and disclosing information from this platform since they have decided together the purpose of the processing and the means to be used (the existing platform). Each of the institutes however is a separate controller for any other processing that may be carried out outside the platform for their respective purposes.*

According to Article 26 of the GDPR, joint controllers should agree (a written **agreement** is recommended, but not necessary) on their respective responsibilities for compliance with the GDPR. Special attention should be paid to agreeing on procedures facilitating the exercise of data subject rights (who is the contact point, who is responsible for providing data subject with information, how the institutions cooperate in processing requests for access, rectification, erasure etc.). The agreement between joint controllers does not bind the data subjects, who can exercise their rights against any of the joint controllers (e.g. any institution participating in a joint project).

Processor – a person or (more often) organization who processes data **on behalf of the controller**, and only on documented instructions therefrom.
For example, a specialized research data archive can process (store) data on behalf of a research institution (the controller).
The same entity can be a processor (acting on behalf of the controller) for some operations, and a controller for other operations on the same dataset (e.g. when an institution that archives a dataset on behalf of a research centre, while also using it for its own research).

If the controller decides to hire a processor ('outsource' some processing operations) it is the controller's responsibility to choose a processor that provides sufficient guarantees of compliance with the GDPR.
The relation between the controller and the processor should always be governed by a detailed **contract** which regulates the obligations of both parties. Several templates for such contracts are available online; one of such templates (in German) is provided by the Data Protection Authority of Baden-Württemberg.

Data Recipient **–** a person or entity to which personal data are disclosed (communicated). E.g., if institution A sends its research data to institution B, institution B is a recipient. The recipient can be:

- a joint controller, if institutions A and B jointly define how and for what purpose the data will be processed by institution B (e.g. in a joint research project); OR
- a processor, if institution B processes the data exclusively on behalf and under instructions from institution A (e.g., archiving by a specialised institution on behalf of a research centre); OR
- an independent controller, if institution B autonomously decides how and for what purpose it will process the data (e.g., data from a publisher are analysed for research purposes by a research centre).

NOTE: The abovementioned roles (controller, joint controller, processor) are attributed **per purpose** rather than per dataset**.** One dataset can have many controllers, as it can be used for various purposes. Likewise, one entity can act as controller of a dataset for one purpose, and as a processor for another purpose (e.g. an archiving institution that also conducts its own research).
It is useful to always start any GDPR compliance assessment with defining a purpose, like in the table suggested in the handout on accountability.

Data Protection Officer (DPO) – a person within each institution (usually a trained lawyer or information scientists specialized in data protection), whose primary task is to inform and advise the institution and its employees on data protection issues, and to cooperate with the supervisory authority authority.
Make sure you know the Data Protection Officer in your institution, he or she should be your first contact point for any questions regarding personal data.

Data Protection Authority (a.k.a. Supervisory Authority) – an administrative body that supervises data processing, publishes guidelines and can issue sanctions and/or fines. In Germany, there is a Data Protection Supervisory Authority in every Land (which collaborate within the framework of the Data Protection Conference, *Datenschutzkonferenz*); in unitary states, there is one such body per country.

GDPR and Research – the requirement for *appropriate safeguards*

The GDPR also applies to data processing for research and archiving purposes (regardless of whether it is commercial or not). However, many GDPR requirements are alleviated when the processing for scientific research purposes is accompanied with **"appropriate safeguards for the rights and freedoms of data subjects".**

These safeguards can be of technical or organizational nature and they should ensure the respect of the data protection principles (in particular: of data minimization).

The safeguards should be adapted to the type of data and the purposes for which they are processed. They may include, for **EXAMPLE** (see Article 89(1) of the GDPR and §22(2) BDSG):

- pseudonymization;
- anonymization as soon as possible (i.e., as soon as the research purposes do not require identifying information to be processed);
- specific training of staff that processes personal data to increase their awareness of data protection principles and risks associated to personal data processing;
- data encryption;
- strict access control (e.g., data can only be accessed by a small group of authorized researchers);
- system of logs that enables to track any modifications (input, alternation, deletion) to the data and attribute them to a specific user;
- processing data in a technically secure and sustainable environment (e.g. where it is not possible to download the data, make copies or print screens; also, where access to data can be maintained in case of technical problems, e.g. through backup copies);
- regular tests of the applied technical and organizational measures (e.g. attempting to re-identify the data, access, copy or delete them without authorization, etc.).

Choosing the appropriate safeguards can be a part of a Data Protection Impact Assessment.

Data protection principle 1a: Lawfulness

According to the principle of lawfulness, processing of personal data should have a **legal basis**.

Available Legal bases are **listed** in Article 6(1) of the GDPR (for 'regular' personal data) and in Article 9(2) of the GDPR (for special categories of personal data).

For 'regular' personal data, two legal bases seem available for research:

- **consent** of the data subject (Article 6(1)(a) of the GDPR) and
- **legitimate interest** (Article 6(1)(f) of the GDPR).

NOTE: In some countries, "public interest" (Article 6(1)(e) of the GDPR) is considered an appropriate legal basis for scientific research. However, this does not seem to be the case in Germany, where traditionally consent is preferred as the go-to legal basis.

For special categories of personal data, the legal bases available for research include:

- **explicit consent** of the data subject (Article 9(2)(a) of the GDPR) and
- **research interest that outweighs the interest of the data subject** (§27(1) BDSG, §13(1) LDSG BW – both adopted on the ground of Article 9(2)(j) of the GDPR).

All these legal bases can be divided into two groups: **consent** and **interest**.

In Article 4 of the GDPR consent is defined as *"any freely given, specific, informed and unambiguous indication of the data subject's wishes by which he or she, by a statement or by a clear affirmative action, signifies agreement to the processing of personal data relating to him or her"*. This definition can be analyzed into the following conditions:

- **any (…) indication of wishes** – the form does not matter (writing, recorded speech/video), as long as proof of consent can be recorded;

- **freely given** – the data subject has to have real choice and control; consent cannot be freely given when there is an imbalance of power (e.g. in employment relationship), if it is the condition to access certain services (unless, of course, the processing of personal data is strictly necessary for the functioning of the services), or if refusal to consent will have negative consequences for the data subject.

  WARNING: paying data subjects for consent, or offering other incentives (such as gifts, free lunches, free services etc.) will likely invalidate their consent.

- **specific** – consent should be given for a specific purpose, and be 'granular' to the extent possible (the data subject should be given the possibility to accept some options, e.g. some purposes, and refuse others); consent should **not** be a 'take-it-or-leave-it' bundle (e.g. research in project X + other research projects at institution Y + long-term archiving by institution Z + commercial research by partner companies). Moreover, consent for data processing should be clearly distinguishable from other aspects, e.g., acceptance of the Terms of Service.

  NOTE: In general, the purpose of the processing should be defined narrowly (e.g. limited to a specific project), but in the context of scientific research, some

flexibility is allowed (e.g. processing "for scientific research in the field of linguistics and digital humanities") – see **Recital 33** of the GDPR (note that the European Data Protection Board calls for a "high degree of scrutiny" [Guidelines 05/2020, para. 155-157] in applying this recital, so use it with caution and try to remain reasonably specific).

- **informed** – to give valid consent, the data subject has to be informed at the very least about:
  - the controller's identity;
  - the purpose(s) of the processing;
  - the type of data that will be collected and used (e.g. name, gender, age);
  - the possibility to withdraw consent;
  - IF APPLICABLE, on the risks related to transfer of his data to countries that do not provide for an adequate level of data protection (see below about data transfers).

  Typically, this information is provided in a written consent form, although it can also be provided orally, or in a short video.

NOTE: The list of information to be provided to data subjects to comply with the principle of transparency and the data subject's right to information is in fact considerably longer. It is a good practice to include all the required information in the consent form, unless this makes the form too long and unreadable for the intended audience, in which case the information can be 'layered' – the elements listed above can be included in the consent form, and full information can be provided in an online document (a Privacy Policy) that data subjects receive a link to.

- **unambiguous** – i.e. unequivocal (a clear "yes", rather than "why not" or "yes, but…");

- **…by a statement or a clear affirmative action…** – silence can never be interpreted as consent. However, consent can (unless it refers to special categories of personal data – cf. below) be implied from one's behavior, if it is sufficiently unequivocal (e.g., clicking a button). Continuing to navigate on a website or scrolling down past a cookie banner are not sufficiently clear to be interpreted as valid consent (note that before the GDPR many websites interpreted scrolling down as consent for cookies; now, clicking a button is required). Accordingly, in online forms boxes signifying consent for data processing should not be pre-ticked (as per the requirement of **Privacy by Default**).

When it comes to the processing **special categories of personal data**, a stricter standard applies – consent for the processing of such data has to be **explicit**. Therefore, it can only be given by an **explicit statement**, and not implied from one's actions. In practice, the difference is rather cosmetic, as data subjects are often asked to signify their consent by an explicit statement, regardless of whether it concerns 'regular' or 'sensitive' data.

Consent can be withdrawn by the data subject at any time, at no cost (Article 7(3) of the GDPR). It should be as easy to withdraw consent as it was to give it (e.g., if consent was given by ticking a box, withdrawal should be as simple as unticking it (one click) – there should be no need to write e-mails or make phone calls).
Withdrawal of consent is not retroactive, i.e., it only produces effects "for the future" – the processing operations that took place before the withdrawal do not become unlawful, but after the withdrawal the processing has to stop, unless a different legal basis (e.g., legitimate interest) applies. If there is an alternative legal basis, the data subject must be duly informed about it according to the transparency principle.

For a detailed analysis of all issues related to consent under the GDPR, see the EDPB guidelines 05/2020.

Interest can be an alternative to consent (although there is no hierarchy between legal bases in the GDPR – they all are "equally good" – the German research community traditionally privileges consent as the "go to" legal basis). Several types of interest can constitute a legal basis for the processing of personal data under the GDPR:

- Legitimate interest (legal basis for the processing of "regular" personal data, Art. 6(1)(f) of the GDPR) of the controller or a third party (including, e.g., the interest of the whole language science community) which is NOT overridden by interests, rights and freedoms of the data subject. In other words, the legitimate benefits that the controller (or a third party) can derive from the processing need to be greater than the risks for the data subject. Moreover, the processing of personal data has to be **necessary** to achieve the purpose of the processing.
  This requires a careful **assessment** (called LIA: legitimate interest assessment), which should take into account the reasonable expectations of the data subject and his relationship with the controller. According to the guidelines of the UK Data Protection Authority (ICO), this assessment should consist of three stages:
  - The **purpose test** (identify the legitimate interest you pursue);
  - The **necessity test** (is the processing of personal data necessary to achieve the purpose?);
  - The **balancing test** (compare the legitimate interest with the data subject's interest).

  EXAMPLE: You decide to store e-mail addresses of people who participated in your speech data collection project, without any link to the data that each participant provided. Your legitimate interest (and the interest of your community) is that when you carry out other similar projects, you can contact the same persons and ask them to participate, as they are likely to agree, which will save you time (and money). E-mail address is the minimum information necessary to easily contact a person. If the addresses are stored securely, in a password-protected environment, the risk for data subjects is minimal (at the very worst, they will receive an unwanted e-mail). Therefore, legitimate interest is an appropriate legal basis.

  The data subject can object to the processing based on legitimate interest "on the grounds related to his or her personal situation".
  For more information about legitimate interest, see the WP29 guidelines 06/2014.

- **Vital interest** of the data subject or another person where the data subject is incapable of giving consent (for **special categories of personal data**, Article 9(2)(c) of the GDPR) – this lawful basis seems specific for medical procedures, and rather unavailable for scientific research.
- German legislators (allowed to do so by the GDPR) adopted specific provisions that enable the processing of special categories of personal data for scientific research purposes without consent. Such provisions can be found in §27 of the BDSG (for private research institutions) and, e.g., in §13 of the BW LDSG (for universities in BW). In short, they allow for special categories of personal data to be processed for research purposes if the processing is necessary to achieve the purpose and if the **research interest in the processing substantially outweighs the interest of data subjects** (the test is therefore similar, but stricter than in the case of legitimate interest assessment). Moreover, the controller is required to implement appropriate safeguards for rights and freedom of data subjects.
- In some countries (like Finland or Estonia), **public interest** (i.e., interest based on a specific legal obligation) may be a suitable legal basis for the processing of personal data for research purposes by public research institutions; however, this does not seem to be the case in Germany. Typically, this basis is used e.g. by tax authorities. Public interest has to be **substantial** to be a valid legal basis for the processing of special categories of personal data.

Legal bases for the processing of personal data for research purposes – an overview:

Regular personal data	Special categories of personal data
Consent (Art. 6(1)(a) of the GDPR)	Explicit consent (Art. 9(2)(a) of the GDPR)
Legitimate interest (Art. 6(1)(f) of the GDPR)	Research interest that substantially outweighs the interests of the data subject [German law]
Public interest (some countries, not Germany – Art. 6(1)(e) of the GDPR)	Substantial public interest (Art. 9(2)(g) of the GDPR)

Data protection principle 1b: Fairness

The principle of **fairness** complements the principle of lawfulness. It is not enough for the processing to have a formal legal basis (i.e.: be lawful), it also has to be **fair**, i.e., not causing unfair prejudice to the data subject.

EXAMPLE: A data subject has consented for his multimodal speech data to be processed for research and teaching purposes. However, the use of this data in teaching in a way that highlights the speaker's low social status or lack of formal education would be unfair, even though it is not exactly unlawful.

Data subjects should also be treated fairly when they exercise their rights.

Fairness in data processing does not have a clear definition; rather, it refers to the general concepts of **justice**, **equity**, and **good faith**.

NOTE: Love thy neighbor!

Data protection principle 1c: Transparency

The **transparency** principle requires that certain **information** about the controller and the processing of personal data be made available to the data subject, allowing him or her to make informed decisions on whether he or she wishes to enter into a relationship with the controller, or exercise his or her rights. It is closely related to the principle of fairness.

The principle of transparency is embodied in the rights of data subjects, and in particular the **right to information** and the **right of access**.

Information to be provided to data subjects under the right to information:

<table>
<tr><th>Data is collected directly from the subject</th><th>Data is obtained from a different source</th></tr>
<tr><td colspan="2">Identity of the controller (+ contact details)</td></tr>
<tr><td colspan="2">Contact details of the DPO</td></tr>
<tr><td colspan="2">Purposes of the processing</td></tr>
<tr><td colspan="2">Legal basis
(if legitimate interest – specify the interest;
if consent – inform about the right to withdraw consent)</td></tr>
<tr><td colspan="2">Categories of recipients (if applicable)</td></tr>
<tr><td colspan="2">Intention to transfer data to third countries (if applicable)</td></tr>
<tr><td colspan="2">Retention period, or criteria used to determine it (see above)</td></tr>
<tr><td colspan="2">Existence of the rights of access, rectification, erasure, restriction, portability and the right to object</td></tr>
<tr><td colspan="2">Existence of automated decision-making, including profiling, and – if applicable – the logic involved and the consequences for the data subject</td></tr>
<tr><td colspan="2">The right to lodge a complaint with a Data Protection Authority</td></tr>
<tr><td rowspan="2">Whether provision of personal data is required by law or necessary to enter into a contract, and if not: what are the consequences of failure to provide the data</td><td>Categories of personal data (e.g., name, age, blog posts, e-mail address)</td></tr>
<tr><td>The source of the data (if applicable – its public availability)</td></tr>
</table>

Data protection principle 2: Purpose limitation

According to the purpose limitation, personal data should be *"collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes"*.

This means that:

- the **purpose** for which you process personal data should be **clearly identified before the processing starts** (e.g., *"for research in the field of language science and digital humanities"*; *"for long-term archiving of research data"*);

WARNING: While the purpose has to be **specific**, there is a real danger in describing it too narrowly. For example: *"for research in the field of Austrian dialectology within YourDilect project at the university of Ulm"* definitely meets the criteria of specificity, but may unnecessarily complicate the re-use of the data for other projects (the data subjects would have to be informed about the new purpose). Often, specifying one or several domains of research (*"research in the field language science and digital humanities*") is specific enough (although it depends on how large the field is, e.g., *"research in the field of social science and the humanities"* may not be specific enough).

- this purpose should be **legitimate** (i.e., briefly put, not illegal) and **documented** (notably in the record of data processing activities);
- the data subject should be informed about the purpose (cf. the transparency principle and the right to information);
- If you want to re-use the data for a different purpose than originally defined, this new purpose has to be **compatible** with the original one. If it is not compatible, you will need a new legal basis for the processing (e.g., new consent). In any case, the data subject has to be informed about the new purpose, whether it is compatible or not.

IMPORTANT NOTE: the GDPR specifically provides that:

- scientific research;
- archiving in the public interest; AND
- statistical purposes

Are **ALWAYS compatible** with the initial purpose. This means that data collected for any purpose (e.g. for research in a different domain or a different project, organizing a scientific conference, or even for direct marketing) can be re-used for the abovementioned purposes without a new legal basis. The data subject, however, has to be informed about the new purpose.

This principle, known as "purpose extention" is crucial for the re-use of legacy data. For more detailed analysis of this issue, see WP29 Opinion 03/2013 (pre-GDPR, but still relevant).

Keep in mind that the processing of personal data for the above-mentioned purposes should be accompanied with **appropriate safeguards**.

Data protection principle 3: Data minimization

Data minimization is probably the single **most fundamental** data protection principle (it has to be observed specifically in **research and archiving** contexts, where its respect needs to be guaranteed by appropriate safeguards). At the same time, this principle is also difficult to reconcile with data-intensive research.

According to this principle, personal data should be *"adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed"*.

In other words, you should not process more data than necessary to achieve the intended purpose.

EXAMPLE: When you collect speech data for a research project, the collection of specific information on data subjects, such as their exact birth date, full postal adress or tax ID number exceeds what is necessary for language research.

Arguably, you should also avoid collecting personal data "just in case", because they may serve in the future, for a different project (for example, recording the city district the person comes from where it is not relevant for the questions asked in the research project for which they are collected).
Rather, you should limit the personal data that you collect and process to what is **strictly necessary** to answer the research questions that are relevant for the project.

By the same token, it is also advisable to **delete or anonymize** personal data as soon as their processing is no longer necessary to achieve the research purpose (Article 89(1) of the GDPR).

EXAMPLE: If you collect student essays to be included in a corpus, the names of their authors may appear on the essays. However, this information is normally not necessary for language research purposes, so you should remove it from your corpus.

The data minimization principle is closely related to the requirement of **Data Protection by Design and by Default** (Article 25 of the GDPR), according to which data protection principles should be "baked into" interfaces and data workflows, and not barely "sprinkled onto" them. For more details about Data Protection by Design in Language Resources, see this LREC 2020 paper by Kamocki and Witt.

Data protection principle 4: Accuracy

According to the principle of accuracy, personal data should be **accurate and, when necessary, kept up to date**.

Arguably, accuracy **only applies to objective facts** (date of birth, surname, e-mail address), and not subjective opinions.

EXAMPLE: If during a recorded interview the interviewee says that in his opinion Lukas Podolski is the best footballer of all time, you will not be obliged to delete or alter this statement just because it is (quite) objectively not true, or when the interviewee changes his opinion. This sort of information (X thinks Y) cannot be objectively verified, and therefore be accurate or inaccurate.
However, if the interviewee says that he grew up in Frankfurt am Oder, and this statement is mistakenly written down as concerning Frankfurt am Main, the information is inaccurate and if this inaccuracy is detected (by a researcher or the data subject himself), it should be rectified.

Historical records remain accurate even if the situation that they concern has changed, as long as it is clear that they are in fact historical records. For this reason, the practical impact of this data protection principle on scientific research is **limited** – for researchers, accuracy is a fundamental quality of any research data.

EXAMPLE: Your speech data indicate the age of the speaker as "35-39 years old" (because the person who collected it was aware of the data minimization principle and decided that recording the exact age of participants would be excessive for the envisaged research purposes). It is obvious that this information is about the age of the speaker at the moment of the recording (it is a historical record), and it is not required to keep it up to date (e.g. by modifying this information every year), which would also be highly counterproductive.

This principle is closely linked to the **right of rectification**.

Data protection principle 5: Storage limitation

According to the storage limitation principle, personal data should be **kept for no longer than necessary** for the purposes for which they are processed.

There are many detailed, country- and sector-specific rules and guidelines on "data retention periods" – maximal (as well as minimal) periods of time for which various documents containing personal data (e.g. CVs, payment slips, data about days off from work, data from security cameras…) can be stored and archived. Defining these retention periods and embedding them in the data workflows is probably one of the biggest data protection challenges for many companies and institutions.

However, where personal data are processed for:
- **Scientific research purposes**
- **Archiving purposes in the public interest OR**
- **Statistical purposes**

With appropriate safeguards, they may be stored for longer periods of time then necessary (as expressly stated in Article 5(1)(e) of the GDPR).

This does not mean, however, that you should keep personal data *ad calendas grecas*. Instead, your **data management plan should include a clearly defined period** after which the data will be either deleted (if they turn out to be of little value) or transferred to a specialized archive for long-time storage. It should be thoroughly examined if it is necessary to store the data in non-anonymized form, or if they can be anonymized before archiving.

While **archived**, the data should also be **periodically reviewed** to evaluate, taking into account the technological progress, e.g.:
- if the data security principle is still observed,
- if the safeguards for data subject rights and interests that are in place are still appropriate, OR
- whether anonymization is possible, or if the data had been anonymized, if the data subjects cannot be identified by means reasonably likely to be used.

Your institution or community may have issued detailed guidelines concerning storage and archiving of personal data for research purposes.

NOTE: The DFG Guidelines for Safeguarding Good Research Practice (a.k.a. Code of Conduct) state the following (in Explanations to Guideline 17: Archiving):

*When scientific and academic findings are made publicly available, the research data (generally raw data) on which they are based are generally archived in an accessible and identifiable manner for a period of **ten years** at the institution where the data were produced or in cross-location repositories. This practice may differ depending on the subject area. In justified cases, **shorter archiving periods may be appropriate**; the reasons for this are described clearly and comprehensibly. The archiving period begins on the date when the results are made publicly available.*

Data protection principle 6: Integrity and confidentiality (a.k.a. data security)

According to the principle of data security (also known as the principle of integrity and confidentiality), personal data should be *"processed in a manner that ensures appropriate **security** of the personal data, including protection against **unauthorized** or unlawful **processing** and against accidental **loss**, **destruction** or **damage**"*. The respect of this principle should be guaranteed by ***"appropriate technical or organizational measures".***

NOTE: These *"appropriate technical and organisational measures"* are distinct from *"appropriate safeguards for the rights and freedoms of the data subjects"* required for research – the former are meant to protect integrity, confidentiality and availability of the data, the latter: all data protection principles and rights of data subjects. The security-related measures can therefore be seen as a sub-group of "appropriate safeguards", but a sub-group that should not be overlooked.

Achieving compliance with the data security principle starts with a **Risk Assessment**. The following table can be used for this (Severtity and Likelihood are evaluated on the scale: 1 – negligible; 2 – moderate; 3 – significant; 4 – maximal). [source: The CNIL guide on data security, p. 4].

Risks	Effects on individuals	Main sources of risks	Main threats	Existing or planned measures	Severity	Likelihood
Illegitimate access to data	Breach of privacy, loss of trust	Lost USB stick	Inappropriate use of data	Password-protection	2	2
Unwanted modification of data	Breach of privacy, loss of trust	Hacked accounts	Loss of data accuracy	Password-protection, data logs	3	2
Loss of data	Loss of trust, Impact on rights	Technical failure	Need to collect new data	Backup copy	3	1

The measures that can be implemented to mitigate the various threats include:

1. pseudonymisation and encryption of personal data;
2. ability to ensure the ongoing confidentiality, integrity, availability and resilience of processing systems and services (e.g. through password-protection, data logs…);
3. ability to restore the availability and access to personal data in a timely manner in the event of a physical or technical incident (e.g. through backup copies);
4. a process for regularly testing, assessing, and evaluating the effectiveness of technical and organizational measures for ensuring the security of the processing (periodic security audits).

Carefully identify the risks that any breach of your research data could entail for data subjects (in language research, these will probably be high only in rather exceptional circumstances, but a lot can depend on the subject of your research), and if you feel that the means at your disposal may not be enough to guarantee appropriate security, get in touch with the DPO at

your institution. Some issues (like securing workstations or the internal network) can only be handled properly at the **institutional level.**

WARNING: Handle with Care, Contains Personal Data! Don't let personal data sit on a USB stick at the bottom of your desk drawer, consider if it is necessary for everyone on your team to have access to the personal data (or is it possible to create a pseudonymized dataset?), make backup copies of your interviews…

Some further guidance on various aspects of personal data security can be find, e.g., in:

- The CNIL (French Data Protection Authority) guide on data security (2018, in English);
- Baden Württemberg Data Protection Authority guidelines on secure passwords (2019, in German).

Proper handling of **data breaches** is also part of ensuring data security.

A data breach (defined broadly as "*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"* [Art. 4 GDPR]) should be **notified** (typically via an online form, if available) to the **Data Protection Authority** (e.g. Der Landesbeauftragte für Datenschutz und Informationsfreiheit Baden-Württemberg) no later than **72 hours** after the data controller becomes aware of it.

IMPORTANT NOTE: The notification obligation **does not apply** when the breach is **unlikely** to result in a risk to the rights and freedoms of natural persons.

In addition, data breaches that are likely to result in a **high risk** to the rights and freedoms of natural persons (which in language research is rather hypothetical), should be **communicated to the data subjects** without undue delay.

All data breaches (even those that do not need to be notified or communicated) should be **documented**. This documentation (which can be included in a specific registry of data breaches) should comprise:

- any **facts** related to the breach;
- its **impact** (when there is no risk for data subjects, this should clearly be mentioned and justified);
- the **remedial action** taken (internal decision-making process, notification/communication).

NOTE: If you discover a possible breach of your research data, make your DPO (or another designated person at your institution) aware of this, so that they can take further steps.

For more information about the steps to follow in case of **data breaches**, see:

- WP29 guidelines on personal data breach notification under the GDPR
- EDPB Guidelines 01/2021 on Examples regarding Data Breach Notification

Data protection principle 7: Accountability

According to the accountability principle, **the controller** is **responsible** for compliance with the GDPR, and should be **able to demonstrate** this compliance.

To be able to achieve and demonstrate compliance with the GDPR, data controllers put in place a number of **measures**, including:

- drafting and adopting **Data Protection Policies** (to comply with the transparency principle and the right to information);
- signing written **contracts** with data processors (if applicable);
- conducting a documented **risk assessment** and implementing **appropriate measures** to comply with the principle of data security;
- documenting consent or other appropriate legal basis (e.g. legitimate interests) to comply with the principle of lawfulness;
- when data are processed for research, archiving or statistical purposes: implementing appropriate safeguards for rights and freedoms of data subjects;
- documenting and – if necessary – reporting (notifying/communicating) data breaches;
- appoint a data protection officer (if applicable);
- handling (or defining procedures for handling) requests related to data subject rights;
- maintaining a **record of data processing activities**;
- if applicable, carrying out a **Data Protection Impact Assessment**.

Record of Data Processing Activities (Article 30 of the GDPR)

A record of data processing activities is a **written document** (typically in XML format) which lists information about all personal data processing operations (with some exceptions which are irrelevant in the context of research) carried out at a given institution.
The record should contain the following information:

- the name and contact details of the **controller** (or joint controllers) and the **data protection officer**;
- the **purposes** of the processing;
- the categories of **data subjects**;
- the **categories** of personal **data**;
- the categories of **recipients**;
- where applicable, **transfers** of personal data to a third country;
- where possible, the envisaged time limits for erasure of the data (**retention periods**);
- a general description of the technical and organizational **security measures**.

There are several **templates** for data processing records (e.g. the XML one proposed by the French Data Protection Authority, considerably more detailed than required by the GDPR, or a simpler one proposed by the German Data Privacy Conference), but since the register of your processing is meant to be integrated into a larger whole (your Institute's or your university's record), you should **follow the methodology adopted at your institution**.

The **"self-assessment" chart** below contains most of the information (apart from the name of the DPO, which should be known to you) that should be included in the record (and more), **but it is not meant to be a template for a record**. Rather, it is intended to help you assess the compatibility of your processing with the GDPR, and gather all the information that your administration can ask you to provide, so that they can include your processing in the "central" record.

Purpose	Legal basis	Data categories	Data subjects	Controller	Processor	Recipients	Transfers to 3rd countries	Security	Safeguards	Retention period
research	Legitimate interest	essays	Students of X	University of X	no	Researchers at Institute of X, archive	no	Password, pseudonym-isation	Pseudonym-isation	10 years
Operating a helpdesk	consent	Name, e-mail, question, IP address	Users of X	University of X jointly with University of Y	University of Z	Designated experts	no	Authentication procedure	-	2 years

Data **processors** are also obliged to keep a record of processing activities (the content of which is slightly different, see Art. 30(2) GDPR).

The record of data processing activities should be **made available to the Data Protection Authority** on request. It is likely to be the first element controlled during a GDPR audit.

For more information about the record, see the **Guidelines** from the German Data Protection Conference (German only), or from the French Data Protection Authority (in English and in French).

Data Protection Impact Assessment (DPIA) – Art. 35 of the GDPR

Data Protection Impact Assessment (DPIA) is **mandatory** for certain types of personal data processing that may represent a **high risk** for the data subjects. The list of such processing operations is determined for **each country individually**, following some guidelines in the GDPR.

For **Germany**, the list of processing operations for which a DPIA is required was adopted by the **German Data Protection Conference** (EN, DE). The specific types of processing on this list that MAY be relevant for 'ordinary' language research **include**:

No. 9: Use of artificial intelligence to process personal data to control interaction with the data subject or to evaluate personal aspects of the data subject

EXAMPLE: Your purpose is to automatically evaluate the mood of your research participants or their background, based on their body language or linguistic expression, using AI methods (e.g. neural networks).

No. 11: Automated evaluation of video or audio recordings to evaluate the personality of those affected

EXAMPLE: Your purpose is to automatically evaluate the mood of subjects participating in a phone call.

No. 14: Anonymisation of [special categories of personal data], *not only in individual cases (in relation to the number of data subjects and the information per data subject) for the purpose of transmission to third parties*

EXAMPLE: You have a large collection of health data (e.g. a corpus of disordered speech). You intend to anonymize it in order to be able to send it to your partners in the US. You have to carry a DPIA first.

When you come up with a novel idea (e.g., a rating portal for "voice data donors" – cf. item no. 5 on the list adopted by the German Data Protection Conference), it is useful to check the list to determine if you have to carry out a DPIA.

Even where DPIA is not mandatory, it is **advisable** to carry it out, e.g., as an **additional safeguard** for the rights and freedoms of data subjects.

The DPIA should be carried out BEFORE the actual processing starts (i.e., at the **conception** phase).

The DPIA should be thoroughly **documented**. It should contain at least:

- a description of the envisaged **processing** operations;
- if the processing is carried out on the basis of legitimate interest, a description of this **interest**;
- an assessement of the **necessity** and **proportionality** of the processing operations with regards to the purposes (is this REALLY necessary to process the data that I intend to process? Isn't the processing that I envisage excessive for my purposes? Can I achieve equally good results without processing personal data to the extent I envisage?);
- an assessment of **risks** for the rights and freedoms of data subjects (if things "go wrong" – what "bad" things can happen to data subjects? Can they be labelled as uneducated or untrustworthy? Can their privacy be exposed in any way? Can they become victims of phishing?);
- the **measures** envisaged **to mitigate these risks** (safeguards, security measures, etc.) – how will they reduce the risk?

If the DPIA indicates that without any mitigating measures, the envisaged processing would result in a **high risk** for data subjects, the controller should **consult the Data Protection Authority** (see Article 36).

There are various specific **guidelines** on how to carry out a DPIA; see e.g.:

- **Guidelines on Data Protection Impact Assessment (DPIA)**
- Datenschutzkonferenz Kurzpapier Nr. 5: Datenschutz-Folgenabschätzung nach Art. 35 DS-GVO
- Gonscherowski et al., Durchführung einer Datenschutz-Folgenabschätzung gem. Art. 35 DSGVO auf der methodischen Grundlage eines stan- dardisierten Prozessablaufes mit Rückgriff auf das SDM am Beispiel eines „Pay as you drive"-Verfahrens (V 0.10)

- PIA Software – a very useful tool for DPIA created and maintained by the French Data Protection Authority (CNIL)

Rights of data subjects: general information

The data subject has the **right to request certain actions** from the controller regarding the processing of the data subject's data. In the context of research, many of these rights are limited (can only be exercised under certain conditions) or even completely excluded (see the table below).

EXCEPTION: When it comes to the right of information, the controller has to take an active role and provide the information EVEN IF it is not requested by the data subject.

As a rule, the controller is obliged to **facilitate** the exercise of these rights by data subjects. To this end, it is advised e.g. to create a **specific e-mail address** that data subjects can write to with any requests, and define **internal procedures** to facilitate the handling of requests (if an e-mail is sent to the designated address, who does what when?).

Sending an **e-mail** is the most convenient (in most cases), but not the only way for data subjects to exercise their rights. They can also communicate their requests the controller in **other ways** (e.g., on a phone); such requests should also be answered, preferably *via* the same communication channel.

The requests should be made by the data subject or his or her legal guardian. If you have reasonable doubts concerning the **identity** of the data subject, you can ask him or her for **additional confirmation**.

The requests should be answered in **clear and plain language** (if the context or courtesy justifies it, in the language of the request), and within **one month**. The possible answers fall within one of the three categories:

- *We have acted on your request* (here is what you asked for: …)
- *We need more time to act on your request, because…*
  - this kind of answer has to be justified by either the **number** of requests (*we have received many request*) or the **complexity** of the request (*acting on your request is complicated*);
  - maximum deadline for answering a request is **three months in total**, so this answer gets you no more than two extra months.
- *Thank you for your e-mail, but we will not act on your request*
  - this kind of answer is justified in the following circumstances:
    - the right that the data subjects wants to exercise is **limited** (e.g. in the context of research or archiving);
    - the request is ***manifestly unfounded or excessive*** (only to be used very exceptionally, e.g., if a data subject is really bombarding you with unjustified requests – in such cases, you can either refuse to act OR charge a **reasonable fee**);
    - the relevant data have been **deleted or anonymized** (as it was not/no longer necessary for our purposes), so the request cannot be answered (there is no obligation to retain data only to be able to answer potential requests from the data subject – cf. Article 11 and Recital 64 of the GDPR).

- this answer should also indicate that the data subject has the **right to lodge a complaint** with the Data Protection Authority.

The requests should be answered free of charge. Only very exceptionally can the controller charge a **reasonable fee** (considering the administrative cost of handling the request). This is the case when:

- the request is ***manifestly unfounded or excessive*** (e.g. unreasonably repetitive), OR
- the data subject requests **more than one copy** of his or her data under the right of access (the first copy should be free).

GDPR Article	Right	In research	In archiving
12-14	Information	Limited (GDPR)	Limited (GDPR)
15	Access	Limited (German law)	Limited (German law)
16	Rectification	Limited (German law)	Excluded (German law)
17	Right to be forgotten	Limited (GDPR)	Limited (GDPR)
18	Restriction	Limited (German law)	Limited (German law)
20	Portability	Applicable	Limited (German law)
21	Right to object	Limited (German law)	Limited (German law)
22	Freedom from automated decision-making	Applicable	Applicable

Rights of data subjects 1: Right to be provided with information (a.k.a. information right)

The information right (Articles 12-14 of the GDPR) is closely linked to the principle of transparency. The controller has to take an **active role** and provide the information even if it is not requested by the data subject.

Data subjects should be provided with information about the processing, regardless of whether the data are collected **directly from the data subject** (e.g., during an interview) or **from a different source** (e.g., downloaded from the Internet).

<table>
<tr><th>Data is collected directly from the subject</th><th>Data is obtained from a different source</th></tr>
<tr><td colspan="2">Information to be provided</td></tr>
<tr><td colspan="2">Identity of the controller (+ contact details)</td></tr>
<tr><td colspan="2">Contact details of the DPO</td></tr>
<tr><td colspan="2">Purposes of the processing</td></tr>
<tr><td colspan="2">Legal basis
(if legitimate interest – specify the interest;
if consent – inform about the right to withdraw consent)</td></tr>
<tr><td colspan="2">Categories of recipients (if applicable)</td></tr>
<tr><td colspan="2">Intention to transfer data to third countries (if applicable)</td></tr>
<tr><td colspan="2">Retention period, or criteria used to determine it (see: storage limitation)</td></tr>
<tr><td colspan="2">Existence of the rights of access, rectification, erasure, restriction, portability and the right to object (the latter should be mentioned clearly and separately – cf. Art. 21(4) GDPR)</td></tr>
<tr><td colspan="2">Existence of automated decision-making, including profiling, and – if applicable – the logic involved and the consequences for the data subject</td></tr>
<tr><td colspan="2">The right to lodge a complaint with a Data Protection Authority</td></tr>
<tr><td rowspan="2">Whether provision of personal data is required by law or necessary to enter into a contract, and if not: what are the consequences of failure to provide the data</td><td>Categories of personal data (e.g. name, age, blog posts, e-mail address)</td></tr>
<tr><td>The source of the data (if applicable – its public availability)</td></tr>
<tr><td colspan="2">When to provide the information?</td></tr>
<tr><td rowspan="3">At the time of collection</td><td>Within a reasonable period of time, but no later than one month after obtaining the data</td></tr>
<tr><td>If the data are used to contact the data subject (e.g. e-mail address): at the latest at the moment of first contact</td></tr>
<tr><td>If disclosure to other recipient(s) is envisaged: at the latest at the moment of first disclosure</td></tr>
<tr><td colspan="2">Exceptions</td></tr>
<tr><td colspan="2">The data subject already has the information</td></tr>
<tr><td></td><td>the provision of information is impossible or would involve a disproportionate effort. This exception applies in particular to</td></tr>
</table>

	research and archiving. In determining whether this exception can apply, account should be taken of: - **the number of data subjects (the bigger the number, the better);** - **the age of the data (the older the data, the better);** - **appropriate safeguards implemented.** **This exception will often allow controllers to "escape" the obligation to inform data subjects in scenarios involving re-use of legacy data.** **However, when the exception applies the information about the processing should be made publicly available.**
	Provision of information would render impossible or seriously impair the objectives of the processing (incl. research)
	The obtention or disclosure of the data result from a legal obligation
	The data is subject to a legal obligation of secrecy

The information is usually provided in one document referred to as **Privacy Policy** (or, in some contexts, directly in **the consent form**).

For a detailed analysis of the principle of transparency and the information right, see WP29 Guidelines on Transparency.

For an online **tool** that will help you generate a GDPR-compliant Privacy Policy for your research project or other research-related activities, see the DARIAH ELDAH Consent Form Wizard.

Rights of data subjects 2: Access

The right of access (Article 15 GDPR), closely linked to the principle of transparency, enables the data subject to **request**:

- a **confirmation** as to whether his or her personal data are being processed, and if this is the case:
- **information** about the processing (most of the information listed above in the right of information); AND
- a **copy** of his or her data that are undergoing processing
  - the first copy should be **free of charge**, a reasonable fee (based on administrative costs) may be charged for further analogue copies;
  - if the request was made by e-mail, the copy should be sent in a **commonly used electronic forma**t;
  - the copy should not contain personal data of other data subjects.

Access is "the mother of all rights", as without it the data subject is unlikely to exercise his other rights (such as rectification, erasure etc.).

The GDPR allows national legislators to limit this right in the context of research (Article 89(2) GDPR) and archiving (Article 89(3) GDPR).

Under **German** law, in the context of **research** (with appropriate safeguards), the right of access is **limited** (cannot be exercised):

- If it is likely to ***render impossible or seriously impair*** the achievement of the research purposes (cf. 27(2) BDSG (Federal) or 13(4) LDSG BW); OR
- If the provision of information requires **disproportionate effort** (*idem*).

Still under **German** law, in the context of **archiving** (with appropriate safeguards), the right of access is **limited** (cannot be exercised) if the data subject's **name does not appear** in the archival material and no information is given which would enable the archival material concerning the data subject to be found with reasonable administrative effort (cf. 28(2) BDSG (Federal) or 14(2) LDSG).

Therefore, in both research and archiving contexts it is **possible to decline an access request** from a data subject, provided that the controller can demonstrate that the abovementioned conditions are met.

Rights of data subjects 3: Rectification

Closely linked to the principle of accuracy, the right of rectification (Article 16 GDPR) enables the data subject to **request** that his or her data be **rectified** (corrected) or **completed**, if they are inaccurate or incomplete (considering the purpose of the processing).

EXAMPLE: Your institution is sending a newsletter to researchers; every e-mail starts with a greeting "Dear [Title] [Surname]". Due to an error in data entry, e-mails sent to Professor Dr Schmitt startswith "Dear Dr. Beckenbauer". Professor Dr. Schmitt can send a request to have his name rectified and his title completed.

Only **objectively verifiable** information can be rectified (dates, names, etc.); statements such as opinions or assessments are beyond the scope of the right to rectification. However, rectifying such statements may be a matter of **good scientific practice**.

EXAMPLE: A speaker from Dortmund says in an interview that Schalke is his favorite football club. When he sees the transcription of his interview, he requires this information to be rectified (his favorite club is of course the BVB) or at least completed (marked as irony), arguing that it was clear from the context (he was wearing a BVB jersey) and his body language that he said it ironically. Arguably, this is beyond the right of rectification, but correcting this statement might be a matter of good practice.

If the data were **disclosed** to any recipients (e.g., shared with a project partner), after a successful rectification request the controller should **notify** the rectification to each of those recipients, **unless** this proves **impossible** or requires **disproportionate effort**. The data subject can request from the controller a list of all the recipients.

The GDPR allows national legislators to **limit** this right in the context of research (Article 89(2) GDPR) and archiving (Article 89(3) GDPR).

Under **German** law, in the context of **research** (with appropriate safeguards), the right of rectification is **limited** (cannot be exercised) if it is likely to ***render impossible or seriously impair*** the achievement of the research purposes (cf. 27(2) BDSG (Federal) or 13(4) LDSG BW).

Still under **German** law, in the context of **archiving** (with appropriate safeguards), the right of rectification is **in principle excluded** (cannot be exercised). However, if the data subject disputes the accuracy of his or her personal data, **he or she can present his or her version**, which shall be added to the archived files (cf. 28(3) BDSG (Federal) or 14(2) LDSG).

Therefore, in both research and archiving contexts it is **possible to decline a rectification request** from a data subject, provided that the controller can demonstrate that the abovementioned conditions are met.

Rights of data subjects 4: Erasure (a.k.a. the right to be forgotten)

The right to be forgotten (Article 17 GDPR) enables the data subject to **request** the controller to **erase** the data subject's data if:

- they are **no longer necessary** for the purpose for which they are processed; OR
- processing of the data is based on **consent** and the data subject has **withdrawn** it (and there is no other legal basis available); OR
- processing is based on legitimate or public **interest** to which the data subject has successfully **objected** (by exercising the right to object); OR
- the data have been processed without a legal basis (i.e., against the principle of lawfulness); OR
- there is a legal **obligation** on the controller to erase the data; OR
- the data were collected from a **minor** *via* an **online service** (e.g., Facebook).

If the data have already been **made public** (e.g., online), the controller should take **reasonable steps** to have any copies of the data or the links to them deleted. Moreover, the controller should **notify** the erasure to all the recipients of the affected dataset, **unless** this proves **impossible** or involves **disproportionate effort**. The data subject can request from the controller a list of all the recipients.

This right can (theoretically) be used to request erasure, e.g., of compromising images or videos (processed without a legal basis) from the Web, but its effects may be **limited** (because it is quite impossible to have all the copies deleted from the Internet, even if far-reaching (beyond reasonable) steps are taken).

The right to erasure is **limited** when the data are processed for **research** or **archiving** purposes (with appropriate safeguards) if the exercise of this right is ***likely to render impossible or seriously impair*** these purposes.

Therefore, in both research and archiving contexts it is **possible to decline an erasure (right to be forgotten) request** from a data subject, provided that the controller can demonstrate that the abovementioned condition is met.

Rights of data subjects 5: Restriction of processing

The right of restriction (Article 18 GDPR) enables the data subject to request **restriction** ("blocking") of his or her data if:

- the **accuracy** of the data is being **disputed** under the right of rectification; OR
- the data is processed **without a legal basis** (against the principle of lawfulness); OR
- the data is **no longer necessary** for the purpose for which it is processed; OR
- the data subject has **objected** (exercised the right to object) to the processing based on legitimate or public interest, and his objection is being **examined** by the controller.

Restricted ("blocked") data are **not deleted**, but they **can only be processed**:

- with **consent** of the data subject; OR
- for exercising **legal claims** (e.g., in court); OR
- for **protecting** other persons; OR
- for reasons of **important public interest**.

EXAMPLE: X discovers his indecent pictures are available on a popular website. X suspects that they were uploaded by Y without X's consent and intends to sue Y for damages. X is considering asking the operator of the website (the data controller) to delete the pictures under the right of erasure, but then realizes that he will need them as a proof. He decides to request restriction instead – the pictures can no longer be shown on the website, but they are still present on the server and can be used for exercising his claim in court.

If the data were **disclosed** to any recipients (e.g., shared with a project partner), after a successful restriction request the controller should **notify** the restriction to each of those recipients, **unless** this proves **impossible** or requires **disproportionate effort**. The data subject can request from the controller a list of all the recipients.

The GDPR allows national legislators to **limit** this right in the context of research (Article 89(2) GDPR) and archiving (Article 89(3) GDPR).

Under **German** law, in the context of **research** and **archiving** (with appropriate safeguards), the right of restriction is **limited** (cannot be exercised) if it is ***likely to render impossible or seriously impair*** the achievement of the purposes (for research, cf. 27(2) BDSG (Federal) or 13(4) LDSG BW; for archiving, cf. 28(4) BDSG (Federal) or 14(4) LDSG BW).

Therefore, in both research and archiving contexts it is **possible to decline a restriction request** from a data subject, provided that the controller can demonstrate that the abovementioned condition is met.

Rights of data subjects 6: Data portability

The right of portability (Article 20 GDPR) enables the data subject to **request**:

- a **copy** of the data that he provided to the controller in a **structured, commonly used and machine-readable format** AND
- a **transmission** of the copy directly to **another controller**, if technically feasible (the data subject can also transmit the data to another controller himself).

This right **only applies** to data that are processed:

- on the basis of **consent** or a **contract** AND
- by **automated means** (not manually).

NOTE: A portability request does not in itself prevent the original controller from continuing to process the data.

EXAMPLE: An individual wants to change his travel agency. Under the portability right, he may ask his current agency to transmit a copy of his personal data (e-mail, phone number, address, banking details, travel preferences, travel history…) to a new agency, to facilitate the transition.

The right of portability was designed to **promote competition** between service providers on the EU single market. Its impact on academia, based on our experience so far, remains hypothetical, although theoretically it could be applied, e.g., to migrate one's data (contact details, papers, connections, preferences) from one repository to another.

The GDPR does not include any exceptions from this right in the context of research. However, national legislators are allowed to **limit** this right in the context of archiving (Article 89(3) of the GDPR).

Under **German** law, in the context of **archiving** (with appropriate safeguards), the right of portability is **limited** (cannot be exercised) if it is likely to ***render impossible or seriously impair*** the achievement of the archiving purposes (28(4) BDSG (Federal) or 14(4) LDSG BW).

For more information on data portability, see WP29 guidelines on the subject.

Rights of data subjects 7: Right to Object

Where personal data are processed on the basis of legitimate or public **interest** (Articles 6(1)(e) and (f) of the GDPR), the right to object (Article 21 GDPR) enables the data subject to **object** to the processing **at any time, *"on the grounds related to his or her particular situation"*** (i.e.: for any reason). This right can be seen as an equivalent of withdrawal of consent for processing based on interest.

The controller should then **stop** processing the data (other than necessary for exercise of legal claims), **unless** he can demonstrate ***compelling legitimate grounds*** for the processing which **override** the interests, rights and freedoms of the data subject.

In other words, the data subject's objection can be overridden if the controller demonstrates *compelling* reasons for that – which, logically, is stricter than simple *legitimate interest*. Currently, in the absence of (long-awaited) guidelines on this right, little is known for sure about the standards that these *compelling legitimate grounds* must meet.

EXCEPTION: When data are processed for **direct marketing** purposes, the data subject can **always** object to it, and the controller has no other option but to stop processing the data.

The GDPR allows national legislators to **limit** this right in the context of research (Article 89(2) GDPR) and archiving (Article 89(3) GDPR).

Under **German** law, in the context of **research** and **archiving** (with appropriate safeguards), the right to object is **limited** (cannot be exercised) if it is ***likely to render impossible or seriously impair*** the achievement of the purposes (for research, cf. 27(2) BDSG (Federal) or 13(4) LDSG BW; for archiving, cf. 28(4) BDSG (Federal) or 14(4) LDSG BW).

Therefore, in both research and archiving contexts it is **possible to override** a data subject's right to object (i.e. to decline a request from a data subject), provided that the controller can demonstrate that the abovementioned condition is met. Furthermore, in our opinion the social benefits derived from research and archiving will often constitute *compelling legitimate grounds* for the processing.

On the other hand, allowing the data subjects to **"opt out"** and withdraw their data, e.g., from a language resource, even where this is not required under the right to object, can be regarded as a **good practice** and a safeguard for data subjects' rights and freedoms.

Rights of data subjects 8: Freedom from automated decision-making and profiling

The right provided for in Article 22 of the GDPR is in fact a general **prohibition** to make **decisions** which produce **legal effects** or similarly **affect** the data subjects (e.g., in job recruitment or in credit applications) **solely by automated means** (i.e., without any human intervention). This prohibition also applies to **profiling** (automated processing of personal data to evaluate, analyze or predict the behavior or the situation of the data subject). Such processing is only allowed in very exceptional circumstances.

NOTE: Arguably, in language research "decisions" that can be made by automated means (e.g. recommending a specific wording for a translation, or suggesting a language resource or a tool) normally do not qualify as *"decisions producing legal effect or similarly affecting the data subject".* Therefore, they are beyond the scope of this prohibition.

This "right" is **not limited** in the context of research or archiving, but its practical impact remains very narrow: it does not apply to situations where an automated decision is subsequently approved by a human being.

For more information about this right, see WP29 Guidelines on the subject.

International Transfers of Personal Data

Movement of personal data within the **European Economic Area (EEA)** is **not restricted**. Between the EEA countries, personal data can be transferred freely, providing that all the GDPR principles (pay special attention to purpose limitation and transparency in this context!) are observed.

NOTE: EEA includes all the EU countries, Iceland, Lichtenstein and Norway. Currently (as of 2021) Switzerland and the UK are not members of the EEA.

Transfers **outside the EEA** are allowed:

- if the 'importing' country has been formally recognized as providing an **adequate level of data protection** (Article 45 GDPR);

  NOTE: So far, the following countries have been recognized by the European Commission as providing adequate level of data protection: Andorra, Argentina, Canada, Faroe Islands, Guernsey, Israel, Isle of Man, Japan, Jersey, New Zealand, Switzerland, United Kingdom and Uruguay. The procedure for recognizing South Korea as one of such countries has been launched in June 2021.

  NOTE: The US are currently **not** recognized as providing an equivalent level of data protection.

  OR ELSE:

- if the transfer is subject to appropriate **safeguards** provided for by **a legally binding instrument**, i.e.:
  - approved **Binding Corporate Rules** (Article 47 GDPR) OR
  - an *ad hoc* **contract** containing **standard data protection clauses** published by the European Commission (SCC) OR
  - an approved **GDPR Code of Conduct** (pursuant Article 40 GDPR);

  OR ELSE

- **exceptionally**, if one of the conditions of Article 49 of the GDPR is met, e.g.:
  - the data subject has given his **explicit consent** for the transfer, after having been **informed** of the associated risks OR
  - the transfer is necessary for **important reasons of public interest**.

NOTE: For many years, data transfers to the US were made possible by the EU-US Privacy Shield agreement (parties to this agreement were recognized as providing for an adequate level of protection). However, the CJEU invalidated this framework in July 2020 (case C-311/18, *Schrems 2*). As of mid-2021, the negotiations on the arrangement are ongoing.

You should generally **refrain** from transferring personal (i.e. non-anonymised) data outside the EEA; however, if you intend to do so, you should first **contact the DPO** at your institution and possibly consider getting professional legal advice.