SEPTEMBER 2018

# LANGUAGE RESOURCES AND RESEARCH UNDER THE GENERAL DATA PROTECTION REGULATION

Pawel Kamocki, Erik Ketzan, Julia Wildgans

Distributed by the CLARIN Legal Issues Committee (CLIC), http://clarin.eu

## Summary

The General Data Protection Regulation (GDPR) will become applicable on 25 May 2018 and repeal the Personal Data Directive of 24 October 1995. It will apply uniformly in all the EU Member States, without need for transposition in national law.

The GDPR builds on the concepts of the Personal Data Directive (personal data, sensitive data, data subject, controller, processor, etc.). However, some elements (such as express recognition of pseudonymisation or requirement of data protection by design and by default) are new. The main principles of the GDPR include: lawfulness, fairness and transparency, purpose limitation, data minimisation, accuracy, storage limitation, integrity and confidentiality, and accountability.

The GDPR also reinforces the rights of data subjects (such as information, access, rectification, erasure, restriction and objection to the processing) and the obligations of data controllers (e.g. obligation to conduct a Data Protection Impact Assessment) and processors (e.g. obligation to keep a record of processing operations).

Some flexibility is allowed for research activities and for archiving in the public interest. In particular, purpose extension (an exception to the principle of purpose limitation) enables re-use for research purposes of data collected for another purpose. In order not to paralyze research projects, certain rights of data subjects are also limited (or can be limited by national legislators) in the case of research. However, all these exceptions can only apply if “appropriate safeguards” to protect the rights and freedoms of the data subject are implemented. It is not entirely clear what can constitute such “appropriate safeguards” (the GDPR expressly only mentions one of them: pseudonymisation).

The GDPR also promotes bottom-up standardisation through e.g. Codes of Conduct or data security certificates, marks and seals. Such instruments, if approved by the competent authorities, may facilitate the application of the GDPR by clarifying its grey zones. The language research community should be encouraged to take advantage of these new opportunities.

# Introduction

The General Data Protection Regulation (hereinafter: GDPR), EU Regulation 2016/679 of 27 April 2016, will become applicable on 25 May 2018 and repeal the Personal Data Directive of 24 October 1995.

Unlike a directive, which requires transposition into national laws (while leaving the choice of "forms and methods"[1] to the Member States), a regulation is binding and directly applicable in all Member States. This means that when the GDPR becomes applicable, all the EU countries will have the same rules regarding the protection of personal data — at least in principle, since some details (including in the area of research — see below) are expressly left to the discretion of the Member States.

The GDPR is a particularly ambitious piece of legislation (consisting of 99 articles and 173 recitals) whose intended territorial scope extends beyond the borders of the European Union[2]. Its main concepts and principles are essentially similar to those of the Personal Data Directive, but enriched with interpretation developed through the case law of the CJEU and the opinions of the Article 29 Data Protection Working Party[3] (hereinafter: WP29).

This White Paper will discuss the main principles of data protection and their impact on language resources, as well as special rules regarding research under the GDPR and the standardisation mechanisms recognized by the Regulation.

## I. General principles of the GDPR

The GDPR builds on the concepts of the Personal Data Directive. Some elements, however, are new. The following paragraphs will present an overview of the data protection framework under the GDPR.

### A. Basic terminology

The key concepts in the GDPR include:

[1] Art. 288 of the Treaty on the Functioning of the European Union.

[2] Cf. art. 3 of the GDPR, according to which the Regulation also applies to the controllers

[3] WP29 is a is an advisory body made up of a representative from the data protection authority of each EU Member State, the European Data Protection Supervisor and the European Commission (see art. 29 of the Personal Data Directive). Under the GDPR, WP29 will be replaced by the European Data Protection Board (art 68 et seq. of the GDPR).

**a) Personal Data**

Just like the Personal Data Directive, the GDPR defines personal data as "any information relating to an identified or identifiable natural person".[4] This definition is indeed very broad; it can be analysed into four elements:

1) "Any information". Every piece of information, regardless of its form (digital or analog text, sound, image or audiovisual material…) and of its content (facts and opinions, true or false), can potentially constitute personal data[5];
2) "Relating to [a person]". According to WP29,[6] information can relate to a person in three ways:
    - via the content: the information says something about the person;
    - via the purpose: the information can be used to evaluate or influence the status or behaviour of the person (e.g. a call log of a telephone can be used to evaluate the behaviour of its owner);
    - via the result: the information can have an impact on the person's rights and interests (i.e. the person may be treated differently from others; e.g. statistics about the person's performance at work).

3) "Identified or identifiable [person]". A person is identified if he/she is singled out from a group. She is identifiable if she can be identified by any means "reasonably likely to be used"[7] by the controller or by a third person. For example pseudonymised data (see below) are still relating to an identifiable person. In contrast, data that do not concern an identifiable person ("anonymous data") is not personal data and therefore the GDPR does not apply to their processing.

4) "Natural person". Only the information relating to natural (i.e. living) persons is to be considered personal data. This means that the GDPR does not apply to the processing of data concerning deceased people or legal entities — however, such information may often relate (via its purpose or result — see above) to natural persons (e.g. the information about a company may influence the status of its CEO; the in-

[4] Art. 4 no. 1 of the GDPR.

[5] It is worth noting that this element of the definition of personal data is somewhat incoherent. From the point of view of information theory, information is a product of data analysis. Information and data occupy different places in the Knowledge Pyramid (Data-Information-Knowledge), it is therefore erroneous, from this point of view, to define (personal) data as information. It seems that the concept of "information" used in the GDPR is different from the one in information theory — in the GDPR, "information" is a synonym of "data".

[6] Article 29 Data Protection Working Party, Opinion 4/2007 on the concept of personal data (WP136), pp. 9-12.

[7] Recital 26 of the GDPR, which further reads: "To ascertain whether means are reasonably likely to be used to identify the natural person, account should be taken of all objective factors, such as the costs of and the amount of time required for identification, taking into consideration the available technology at the time of the processing and technological developments".

formation about the cause of death of one's ancestor may influence his behaviour etc.).

Moreover, certain categories of personal data listed in art. 9 of the GDPR (and called "special categories of personal data", or "sensitive data") are subject to stricter protection. These special categories of data include (but are not limited to) those revealing a person's racial or ethnic origin, political opinions, religious or philosophical beliefs, health and sex life or sexual orientation.

**b) Processing**

The GDPR defines processing broadly as "any operation or set of operations which is performed on personal data (...), whether or not by automated means"[8]. This includes (but is not limited to) collection, recording, annotation, storage, consultation, making of backup copies, but also anonymisation (the process of making personal data anonymous)[9] or erasure.

**c) Data Subject**

The data subject is the natural person that the personal data relate to. The data subject has certain rights with regards to his data (see below).

**d) Data Controller**

Data controller is "the natural or legal person (...) which, alone or jointly with others, determines the purposes and means of the processing of personal data"[10]. It is therefore the person (e.g. project lead) or entity (e.g. institution or consortium) who decides why and how personal data are to be processed. There can be several controllers for one processing (e.g. an entity that defines the purposes of processing and a technical expert who decides on the means of processing). Since the role of data controller incurs specific obligations and responsibilities (see below), it is useful to clearly identify the controller(s) in a Data Management Plan.

**e) Processor**

Processor is "a natural or legal person (...) which processes personal data on behalf of the controller".[11] In other words, it is a person (e.g. a research assistant) or entity (e.g. an archive) that processes personal data (e.g. stores it) only on instructions of the controller,[12] without making any decisions as to the means and purposes of the processing. A processor

[8] Art. 4, no. 2 of the GDPR.
[9] For further information see: Article 29 Data Protection Working Party, Opinion 05/2014 on Anonymisation Techniques (WP216)
[10] Art. 4 no. 7 of the GDPR.
[11] Art. 4 no. 8 of the GDPR.
[12] Art. 29 of the GDPR.

is an optional part of the processing chain: the controller may choose not to hire a processor and simply carry out the processing by himself.

**f) Data Protection Officer**

A Data Protection Officer (DPO) needs to be appointed by certain categories of data controllers and processors (including universities and most — if not all — research institutions).[13] His tasks include providing controllers and processors with information and advice, monitoring compliance with the GDPR and acting as liaison between his institution and the supervisory authority.[14] All researchers working with personal data are advised to contact the DPO in their institution in order to obtain information and advice.

**g) Pseudonymisation**

Pseudonymisation is defined as "processing of personal data in such a manner that the personal data can no longer be attributed to a specific data subject without the use of additional information, provided that such additional information is kept separately and is subject to technical and organisational measures to ensure that the personal data are not attributed to an identified or identifiable natural person".[15] Despite being a well-known and frequently used (also in research) measure, pseudonymisation was not mentioned in the Personal Data Directive; its introduction in the GDPR can be seen as an important advancement for language resources. It needs to be stressed that from the legal point of view pseudonymisation does not have the same effect as anonymisation (which needs to be irreversible): pseudonymised data are still personal data,[16] and therefore subject to the GDPR rules. Pseudonymisation is merely an expressly recognized safeguard (among other possible safeguards — see below) for the rights and interests of data subjects.[17]

## B. Data protection principles

The principles relating to processing of personal data are defined in art. 5 of the GDPR. Most of these principles existed already in the Personal Data Directive, but the GDPR restates some of them with more clarity.

**a) Lawfulness, fairness and transparency**

In principle, processing of personal data is lawful if it is carried out on the basis of the consent of the data subject.

[13] Art. 37 of the GDPR.
[14] Every EU Member State has a National Data Protection Authority (see: http://ec.europa.eu/justice/data-protection/article-29/structure/data-protection-authorities/index_en.htm).
[15] Art. 4 no. 5 of the GDPR.
[16] Recital 26 of the GDPR.
[17] Recital 28 of the GDPR.

Consent[18] is "any freely given, specific (cf. below), informed and unambiguous indication of the data subject's wishes by which he or she (...) signifies agreement to the processing". Consent can be express (by a written or oral statement, including by electronic means)[19] or implied (by an affirmative action). Silence or inaction (e.g. a pre-ticked box) cannot be interpreted as consent.

Consent can be withdrawn at any moment, but this withdrawal is not retroactive (i.e. it does not affect the lawfulness of the processing prior to the withdrawal).[20]

For processing of data relating to minors under the age of 16, consent needs to be given or authorised by the holder of parental authority. Member States may lower this "age of consent", but not below 13 years of age.[21]

When it comes to processing of special categories of personal data, consent must be explicit (i.e. no implied consent possible).

Exceptionally, processing of personal data can also be carried out without consent, i.e. on the basis of one of alternative grounds for lawfulness listed in art. 6 of the GDPR. From the point of view of research, the most important of these alternative grounds is art. 6 (1) (f) according to which processing is lawful if it is:

> necessary for the purposes of the legitimate interests pursued by the controller or by a third party, except where such interests are overridden by the interests or fundamental rights and freedoms of the data subject

It is therefore a "balance of interests" test: processing is lawful if the legitimate interests (personal, scientific or societal) outweigh those of the data subject in protecting his privacy. This must be evaluated on a case-by-case basis, taking into account such elements as the character of data that are being processed, the reasonable expectations of the data subject and applied safeguards (such as pseudonymisation). In our opinion, in the context of language research, this ground for processing should only be relied on in very special cases, and subject to careful assessment.

Transparency. According to recital 39 of the GDPR, "[t]he principle of transparency requires that any information and communication relating to the processing of those personal data be easily accessible and easy to understand, and that clear and plain language be used" (see below about the right of data subjects to information).

[18] For more information see: Article 29 Data Protection Working Party, Guidelines on Consent under Regulation 2016/679 (WP259).

[19] "If the data subject's consent is to be given following a request by electronic means, the request must be clear, concise and not unnecessarily disruptive to the use of the service for which it is provided" (recital 32 of the GDPR).

[20] Art. 7(3) of the GDPR.

[21] Art. 8 of the GDPR.

Fairness. Fairness refers to the general concept of justice, equity and reasonableness.

**b) Purpose limitation**

According to the principle of purpose limitation personal data shall be "collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes".[22] In other words, the purpose of processing shall be specified before the processing starts and respected throughout the whole personal data lifecycle. Extensions of the once specified purpose are only possible in specific cases (including for research purposes — see below).

**c) Data minimisation**

According to the principle of data minimisation personal data shall be "adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed".[23] This is a general prohibition of processing (including collection and storage, cf. below on storage limitation) of personal data that are not necessary to achieve the purposes of processing. This principle is therefore largely incompatible with data-intensive operations performed on personal data.

**d) Accuracy**

Personal data shall also be "accurate and, where necessary, kept up to date".[24] This principle is closely related to data subjects' rights of access and rectification (see below). It is also compatible with ethical norms and best practices recognized by the research community.

**e) Storage limitation**

According to the principle of storage limitation personal data shall be "kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed".[25] Long-term storage of personal data is possible only in very specific cases (see below).

**f) Integrity and Confidentiality**

Personal data should also be "processed in a manner that ensures appropriate security of the personal data, including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures".[26] Such security measures are always necessary whenever personal data are being processed; moreover, in certain specific cases (see below), a higher standard may apply.

[22] Art. 5(1)(b) of the GDPR.
[23] Art. 5(1)(c) of the GDPR.
[24] Art. 5(1)(d) of the GDPR.
[25] Art. 5(1)(e) of the GDPR.
[26] Art. 5(1)(f) of the GDPR.

It shall also be noted that ensuring integrity of research data is also a requirement of research ethics and deontology.

**g) Accountability**

The accountability principle is a new and important addition in the GDPR. According to art. 5(2) of the GDPR the controller shall be responsible and able to demonstrate compliance with all the principles of the GDPR. In other words, the burden of proof is always on the controller: it is not for the data subject to prove that the principles of the GDPR are infringed, but for the controller to prove that they are respected.

## C. Rights of data subjects

Apart from the right to consent to the processing, to refuse to consent or to withdraw consent, data subjects have other unwaivable rights related to their data which must be taken into account by controllers and processors. These rights include:

**a) Right to information:**

The data subject has the right to be provided certain information about the processing, such as e.g.:

- the identity of the controller,
- the categories of data concerned,
- the purpose of processing,
- the duration of storage,
- the envisaged transfers and the rights of access and rectification (see below).

This information must be provided to the data subject regardless of whether the data were obtained from him[27] or not.[28] According to the principle of transparency, the information should be in a "concise, transparent, intelligible and easily accessible form, using clear and plain language".[29]

**b) Right of access**

Apart from the information mentioned above, the data subject has the right to obtain from the controller confirmation as to whether personal data concerning him are being processed, and if it's the case, access to the data.[30] The controller should use all reasonable

[27] Art. 13 of the GDPR.
[28] Art. 14 of the GDPR.
[29] Art. 12 of the GDPR.
[30] Art. 15 of the GDPR.

measures to verify the identity of a data subject who requests access and, upon verification,[31] provide a copy of the personal data undergoing processing.

**c) Right to rectification**

The data subject has the right to obtain from the controller without undue delay the rectification of inaccurate personal data concerning him or her.[32]

**d) Right to erasure (right to be forgotten)**

The data subject has the right to obtain (without undue delay) from the controller erasure of personal data concerning him,[33] e.g. if:

- the data are no longer necessary in relation to the purposes for which they were collected;
- the data subject withdraws his consent, and there is no other lawful ground for processing (see above about the principle of lawfulness);
- the data have been unlawfully processed (e.g. in violation of the principle of minimisation or integrity and confidentiality).

**e) Right to restriction of processing**

The right of restriction is similar to the right of erasure; in certain cases (when the processing is unlawful or when the controller needs more time to verify the accuracy of data or seek for an alternative ground for processing), the data subject may choose to request restriction[34] of processing instead of rectification or erasure. The restricted data become "blocked for use": they can still be stored by the controller, but can only be processed with the data subject's consent.

**f) Right to object**

In certain specific circumstances,[35] when processing is based on a different ground for lawfulness than consent (see above), the data subject has the right to object to the processing of personal data concerning him "on grounds relating to his or her particular situation". Unlike the right to erasure, this right can be exercised even if the processing is lawful. If the data subject decides to exercise his right to object and the controller wants to continue to process the data, it shall be for the controller to demonstrate his "compelling legitimate interests"[36] which override the interests, rights and freedoms of the data subject (see below about the right to object in the context of research).

[31] Recital 64 of the GDPR.
[32] Art. 16 of the GDPR.
[33] Art. 17 of the GDPR.
[34] "Restriction of processing" is defined as "the marking of stored personal data with the aim of limiting their processing in the future" (art. 4 no. 3 of the GDPR).
[35] Listed in art. 21 of the GDPR.
[36] Recital 69 of the GDPR.

## D. Obligations of data controllers and processors

The GDPR significantly increases the burden on data controllers. Their obligations include:

### a) Data Protection by Design and by Default

The controller is obliged to implement data protection by design and by default, "[t]aking into account the state of the art, the cost of implementation and the nature, scope, context and purposes of processing as well as the risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing".[37] According to the general principle of accountability (see above), he shall also be able to demonstrate compliance with this obligation.

"Data protection by design" signifies that already at the stage of planning (designing) the processing, the controller shall implement "appropriate technical and organisational measures"[38] to ensure respect of the GDPR.

"Data protection by default" signifies "appropriate technical and organisational measures for ensuring that, by default, only personal data which are necessary for each specific purpose of the processing are processed".[39] In particular, such measures shall ensure that personal data are not made openly accessible without the data subject's express intervention.

According to recital 78 of the GDPR, "appropriate technical and organisational measures" may include, apart from internal privacy policies: "minimising the processing of personal data, pseudonymising personal data as soon as possible, transparency with regard to the functions and processing of personal data, enabling the data subject to monitor the data processing, enabling the controller to create and improve security features". The use of privacy enhancing technologies (PET), a Data Management Plan or privacy policies can be added to this list.

### b) Records of data processing activities

In order to be able to demonstrate compliance with the GDPR (see above about the principle of accountability), both the controller and the processor shall keep a record of processing activities carried out under their responsibility. The information that needs to be

[37] Art. 25(1) of the GDPR.
[38] Ibid.
[39] Art. 25(2) of the GDPR.

included in such a record is listed in art. 30 of the GDPR.[40] The obligation to keep such a record do not apply to processors or controllers that are organisations with fewer than 250 employees, with some exceptions: a record needs to be maintained even by smaller organisations if they process personal data regularly or if the processing concerns sensitive data (listed in art. 9 of the GDPR — see above) or may otherwise result in a risk to the rights and freedoms of the data subjects. Researchers are advised to contact the Data Protection Officer at their institution to know if they are concerned by the obligation to keep such a record; even if they are not, keeping such a record may be regarded as an additional safeguard (necessary to benefit from research exceptions — see below). Moreover, keeping such a record may be interesting from the point of view of Open Methodology and reproducibility of research.

### c) Security; notification of data breaches

The GDPR puts particular emphasis on security. Both the controller and the processor are therefore obliged to implement organisational and technical measures to ensure appropriate level of security, "taking into account the state of the art, the costs of implementation and the nature, scope, context and purposes of processing as well as the risk (...) for the rights and freedoms of natural persons".[41] Such measures may include "minimising the pro-

[40] Information to be included in a record kept by the controller:

- the name and contact details of the controller and, where applicable, the joint controller, the controller's representative and the data protection officer;
- the purposes of the processing;
- a description of the categories of data subjects and of the categories of personal data;
- the categories of recipients to whom the personal data have been or will be disclosed including recipients in third countries or international organisations;
- where applicable, transfers of personal data to a third country or an international organisation, including the identification of that third country or international organisation and, in the case of transfers referred to in the second subparagraph of Article 49(1), the documentation of suitable safeguards;
- where possible, the envisaged time limits for erasure of the different categories of data;
- where possible, a general description of the technical and organisational security measures referred to in Article 32(1).

Information to be included in a record kept by a processor:

- the name and contact details of the processor or processors and of each controller on behalf of which the processor is acting, and, where applicable, of the controller's or the processor's representative, and the data protection officer;
- the categories of processing carried out on behalf of each controller;
- where applicable, transfers of personal data to a third country or an international organisation, including the identification of that third country or international organisation and, in the case of transfers referred to in the second subparagraph of Article 49(1), the documentation of suitable safeguards;
- where possible, a general description of the technical and organisational security measures referred to in Article 32(1).

[41] Art. 32 of the GDPR.

cessing of personal data, pseudonymising personal data as soon as possible, transparency with regard to the functions and processing of personal data, enabling the data subject to monitor the data processing, enabling the controller to create and improve security features".[42] It shall be noted that the standard for appropriate level of security will evolve over time, and therefore the technical and organisational measures implemented shall be periodically reviewed.

If a personal data breach occurs,[43] the controller shall notify it without undue delay to the data protection authority,[44] and — if it results in high risk to rights and freedoms of natural persons — communicate it to the affected data subjects.[45]

**d) Data Protection Impact Assessment (DPIA)**

Art. 35 of the GDPR introduces the concept of a Data Protection Impact Assessment (DPIA). It can be defined as a "process for building and demonstrating compliance [with the GDPR]"[46] (see above about the principle of accountability). Despite being an important tool (also to ensure data protection by design and by default — see above), carrying out a DPIA is not always mandatory; rather, it is required only when the processing is "likely to result in high risk to the rights and freedoms of natural persons" (art. 35(1) of the GDPR). The WP29 recommends to carry out a DPIA even when it is not required by law[47]; such a "facultative" DPIA may also be regarded as an additional safeguard (see below about special rules concerning research). Moreover, the assessment of whether it is necessary to conduct a DPIA shall be periodically renewed, as the result is susceptible to change over time (what was not likely to result in high risk today may become so in a couple of years).

It may seem obvious that processing carried out in language research would rarely result in a high risk for the subjects. Indeed, the examples given in art. 35(3) of the GDPR[48]

[42] Recital 78 of the GDPR.

[43] "Personal data breach" is defined 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 no. 12 of the GDPR).

[44] Art. 33 of the GDPR.

[45] Art. 34 of the GDPR.

[46] Article 29 Data Protection Working Party, Guidelines on Data Protection Impact Assessment (DPIA) and determining whether processing is "likely to result in a high risk" for the purposes of Regulation 2016/679, WP248 rev. 01 (4 October 2017), p. 4.

[47] Article 29 Data Protection Working Party, Guidelines on DPIA, p. 8.

[48] A data protection impact assessment (...) shall **in particular** be required in the case of:

*a)* a systematic and extensive evaluation of personal aspects relating to natural persons which is based on automated processing, including profiling, and on which decisions are based that produce legal effects concerning the natural person or similarly significantly affect the natural person;

*b)* processing on a large scale of special categories of data referred to in Article 9(1), or of personal data relating to criminal convictions and offences referred to in Article 10; or

*c)* a systematic monitoring of a publicly accessible area on a large scale.

are far removed from the reality of language research. However, according the guidelines adopted by the WP29, the criteria to be taken into account in evaluating whether processing may be likely to result in a high risk to the rights and freedoms of natural persons include e.g.:

- the character of processed data, i.e. whether sensitive data (see above) or data relating to criminal convictions or offences are processed;
- the scale of processing, i.e. whether processing is carried out "on a large scale"[49];
- matching or combining datasets, e.g. originating from two or more data processing operations performed for different purposes and/or by different controllers in a way that would exceed the reasonable expectations of data subjects;
- processing of data concerning vulnerable persons (such as children, asylum seekers, the ill, the elderly…)
- innovative use applying new technological solutions (e.g. data collected by connected objects).

The assessment of risk must be carried out on a case-by-case basis, but considering the above it seems that certain categories of language research (e.g. in the field of machine translation, disordered speech or language acquisition) may indeed be concerned by the obligation to carry out a DPIA. When in doubt, one should consult the website of the supervisory authority[50] or the Data Protection Officer.

The DPIA may concern a single processing operation or a set of similar processing operations (e.g. a range of similar research projects)[51], also when they are carried out by different controllers.[52] It is therefore possible to have a "common", EU-wide "reference DPIA" for a category of processing operations (e.g. for research on speech recognition, or perhaps even more broadly: in language research involving audiovisual data). In such a case, according to the WP29's guidelines: "a reference DPIA should be shared or made publicly accessible, measures described in the DPIA must be implemented, and a justification for conducting a single DPIA has to be provided".[53]

[49] According to WP29, the following factors should be considered when determining whether the processing is carried out on a large scale: 1) the number of data subjects concerned; 2) the volume of data and/or the range of different data items being processed; 3) the duration, or permanence, of the data processing; 4) the geographical extent of the processing.
[50] Art. 35(4) requires that supervisory authorities establish a list of processing operations that require a DPIA and communicate it to the European Data Protection Board (EDPB, a body created by art. 68 of the GDPR which will replace WP29); some supervisory authorities offer forms and applications that help carry out a DPIA, see e.g. https://www.cnil.fr/en/pia-software-updates-beta-version.
[51] Recital 92 of the GDPR.
[52] Article 29 Data Protection Working Party, Guidelines on DPIA, p. 7.
[53] ibid.

If the DPIA (which must be conducted prior to the processing) indicates that the processing would result in high risk for the data subjects, the controller shall take measures to mitigate this risk or consult (prior to the processing) the supervisory authority.[54]

### e) Data Protection Officer

Under the GDPR, many data controllers (including universities and most — if not all — other research institutions)[55] are required to appoint a Data Protection Officer (DPO). The DPO shall be involved "properly and in a timely manner in all issues which relate to the protection of personal data".[56] He serves as a liaison between the controller and the data protection authority; his tasks include information, advice and monitoring of compliance. In the performance of these tasks, the DPO is bound by secrecy and confidentiality. Researchers working on personal data are advised to regularly contact the DPO at their institution.

## E. Principles regarding cross-border transfer of personal data

The principles regarding cross-border transfer of personal data are similar to those in the Personal Data Directive:

- personal data can be freely transferred within the European Union (providing, of course, that the principles of the GDPR are respected); in other words, personal data may flow within the borders of the EU as easily as within the borders of one Member State;
- transfer to third countries is possible if:
    - the European Commission has decided that the third country ensures an adequate level of data protection ("adequacy decision") OR;
    - the transfer is subject to appropriate safeguards, such as binding corporate rules,[57] model contracts for the transfer of personal data to third countries,[58] OR;
    - exceptionally, if the data subject has explicitly consented to the proposed transfer, after having been informed of the possible risks of such transfers for the data subject due to the absence of an adequacy decision and appropriate safeguards.

[54] Art. 36 of the GDPR.
[55] See art. 37-39 of the GDPR for more information.
[56] Art. 38(1) of the GDPR.
[57] See: https://ec.europa.eu/info/strategy/justice-and-fundamental-rights/data-protection/data-transfers-outside-eu/binding-corporate-rules_en
[58] https://ec.europa.eu/info/law/law-topic/data-protection/data-transfers-outside-eu/model-contracts-transfer-personal-data-third-countries_en

The rules regarding transfer of personal data, together with other principles of the GDPR, make sharing of personal data under Open Data conditions (i.e. accessible and re-usable for everyone and for any purpose) nearly impossible.

**Countries that provide for an adequate level of data protection**

The European Commission has so far recognised Andorra, Argentina, Canada (commercial organisations), Faroe Islands, Guernsey, Israel, Isle of Man, Jersey, New Zealand, Switzerland and Uruguay as providing adequate protection. Adequacy talks are ongoing with Japan and South Korea.

**Special case: The United States**

Transfer of personal data from the EU to the United States is governed by a special framework called the Privacy Shield, an agreement whereby participating companies are deemed as having adequate protection. As of February 2018, Privacy Shield certification is held by Amazon, Dropbox, Microsoft, Google, Facebook, and 2600+ other entities (for a full list, see: https://www.privacyshield.gov/) To transfer data to a US recipient that has not signed up to the Privacy Shield Framework, researchers should contact their data protection officer who may consider alternative arrangements (see above).

## II. Special rules concerning research under the GDPR

**What is "research" under the GDPR?**

Recital 159 of the GDPR defines research very broadly as "including for example technological development and demonstration, fundamental research, applied research and privately funded research". Therefore, the "non-commercial research only" requirement known from copyright exceptions does not apply to the processing of personal data; commercial research or public-private partnerships can also benefit from the special rules in the GDPR. However, the WP29 considers that "the notion [of research] may not be stretched beyond its common meaning and understands that "scientific research" in this context means a research project set up in accordance with relevant sector-related methodological and ethical standards".[59]

[59] Article 29 Data Protection Working Party, Guidelines on Consent under Regulation 2016/679 (WP259), p. 27.

### A. Flexibility concerning specificity of consent to processing for research purposes

According to the general framework of the GDPR consent needs to be specific, i.e. given in relation to a specific purpose. Processing for research purposes is not completely exempted from this requirement, but recital 33 of the GDPR allows for some flexibility. It recognizes that:

> It is often not possible to fully identify the purpose of personal data processing for scientific research purposes at the time of data collection. Therefore, data subjects should be allowed to give their consent to certain areas of scientific research when in keeping with recognised ethical standards for scientific research. Data subjects should have the opportunity to give their consent only to certain areas of research or parts of research projects to the extent allowed by the intended purpose.

WP29 advocates a very strict interpretation of this recital. According to the Working Party, "scientific research projects can only include personal data on the basis of consent if they have a well-described purpose. Where purposes are unclear at the start of a scientific research programme, controllers will have difficulty to pursue the programme in compliance with the GDPR".[60] A more general description of the purposes of processing (which seems allowed by recital 33) is only possible in some very special cases. Moreover, according to WP29, the application of this more flexible approach necessitates appropriate safeguards (see below).

WP29 further suggests that: "[w]hen research purposes cannot be fully specified, a controller must seek other ways to ensure the essence of the consent requirements are served best, for example, to allow data subjects to consent for a research purpose in more general terms, and for specific stages of a research project that are already known to take place at the outset. As the research advances, consent for subsequent steps in the project can be obtained before that next stage begins".[61]

The Working Party also stresses the importance of a research plan (specifying in plain terms the research questions and envisaged methods) that the data subjects should be able to consult before they give their consent to the processing. Such a plan is also useful to demonstrate the compliance with the requirement for informed (and not only specific) consent[62].

[60] Article 29 Data Protection Working Party, Guidelines on Consent (...), p. 27.
[61] Idem, p. 28.
[62] Idem, p. 28-29.

**B. Exceptions to certain GDPR principles subject to "appropriate safeguards"**

Art. 89(1) of the GDPR is the nucleus of the framework concerning research under the GDPR. Interpreted together with other articles, this norm provides exceptions from certain principles of the GDPR on the condition that "appropriate safeguards for the rights and freedoms of data subjects" are applied (see below).

**a) Exception to the principle of purpose limitation (purpose extension)**

The principle of purpose limitation states that personal data shall be "collected for specified, explicit and legitimate purposes and not further processed in a manner that is incompatible with those purposes" (see above). By extension, processing for research purposes (providing that appropriate safeguards are applied) is always to be regarded as a "compatible purpose".[63] This means that data lawfully collected for any purpose can then be reused for research purposes, without the necessity to collect new consent or seek for another ground for lawfulness of processing.

**b) Exception to the principle of storage limitation**

According to the principle of storage limitation personal data shall be "kept in a form which permits identification of data subjects for no longer than is necessary for the purposes for which the personal data are processed" (see above). However, if data are to be processed solely for research purposes *with appropriate safeguards*, they can be stored for a longer period of time.[64]

**c) No exceptions to the principle of data minimisation**

The principle of data minimisation, though incompatible with data-intensive science, applies fully to the processing of data for research purposes. The respect of this principle is even listed as the main function of the "appropriate safeguards" (see below). As a consequence, data processed for research purposes shall be "adequate, relevant and limited to what is necessary in relation to [the research project]"; moreover, art. 89(1) of the GDPR expressly requires that the data processed for research purposes shall be anonymised as soon as possible, if the purposes of the research can be fulfilled in that manner.

[63] Art. 5(1)(b) of the GDPR.
[64] Art. 5(1)(e) of the GDPR.

## C. Exceptions to certain rights of data subjects

The GDPR contains also some exceptions to the rights of data subjects in case of processing of personal data for research purposes. They can be divided into two groups: those that are "mandatory" and those that are "optional".

The "mandatory" exceptions are already a part of the GDPR and will apply uniformly in all the Member States. They do not require any intervention from the national legislator. The "optional" exceptions are simply allowed by the GDPR, but the decision to implement them is in practice left to the Member States.

### a) "Mandatory" exceptions

**Right to information**

The right to information may be limited, but only when the data have not been obtained directly from the data subject (e.g. repurposing of previously collected data allowed by the mechanism of "purpose extension"). In such cases, the right to information does not apply insofar as "the provision of such information proves impossible or would involve a disproportionate effort (...), or in so far as [it] is likely to render impossible or seriously impair the achievement of the [purposes of research".[65] However, the controller shall take "appropriate measures to protect the data subject's rights and freedoms and legitimate interests, including making the information publicly available".[66]

When the data are obtained from the data subject, there is no exception to the right of information.

**Right to erasure**

The right of erasure does not apply to processing that is necessary for research purposes if the exercise of this right "is likely to render impossible or seriously impair" the achievement of the purposes of the research.[67] In other words, when the data are crucial for the project (which is rarely the case of data-intensive research in the field of linguistics and language technology), the data subject cannot exercise the right to erasure.

**Right to object**

The data subject cannot exercise his right to object to the processing of his data for research purposes if "the processing is necessary for the performance of a [research] carried out for reasons of public interest".[68] It seems therefore that research in linguistics and language technology would only very exceptionally qualify for this exception.

---

[65] Art. 14(5)(b) and recital 62 of the GDPR.
[66] Art. 14(5)(b) of the GDPR.
[67] Art. 17(3)(d) of the GDPR.
[68] Art. 21(6) of the GDPR.

**b) "Optional" exceptions**

Due to controversies on whether the European legislator was competent to create research exceptions to certain rights of data subjects, the decision was left to the national legislators of the Member States. According to the article 89(2):

> Union or Member State law may provide for derogations from the rights [of access, rectification, restriction of processing or the right to object] subject to the conditions and safeguards (...) in so far as such rights are likely to render impossible or seriously impair the achievement of the specific purposes, and such derogations are necessary for the fulfilment of those purposes.

The exceptions to these four rights (access, rectification, restriction or processing and right to object) are therefore to be introduced (or not) by the national legislators who will also specify the conditions of their application.

For now, Germany is one of the few countries that adopted such norms. Art. 27(2) of the New Federal Data Protection Law (*BDSG in der Fassung der Bekanntmachung vom 30. Juni 2018*, BDSG n. F.) implements art. 89(2) nearly word for word.[69] Art. 28 contains similar exceptions for archiving in the public interest.

[69] It is interesting — for our German readers — to quote this article in extenso.

(1) Abweichend von Artikel 9 Absatz 1 der Verordnung (EU) 2016/679 ist die Verarbeitung besonderer Kategorien personenbezogener Daten im Sinne des Artikels 9 Absatz 1 der Verordnung (EU) 2016/679 auch ohne Einwilligung für wissenschaftliche oder historische Forschungszwecke oder für statistische Zwecke zulässig, wenn die Verarbeitung zu diesen Zwecken erforderlich ist und die Interessen des Verantwortlichen an der Verarbeitung die Interessen der betroffenen Person an einem Ausschluss der Verarbeitung erheblich überwiegen. Der Verantwortliche sieht angemessene und spezifische Maßnahmen zur Wahrung der Interessen der betroffenen Person gemäß § 22 Absatz 2 Satz 2 vor.

(2) Die in den Artikeln 15, 16, 18 und 21 der Verordnung (EU) 2016/679 vorgesehenen Rechte der betroffenen Person sind insoweit beschränkt, als diese Rechte voraussichtlich die Verwirklichung der Forschungs- oder Statistikzwecke unmöglich machen oder ernsthaft beeinträchtigen und die Beschränkung für die Erfüllung der Forschungs- oder Statistikzwecke notwendig ist. Das Recht auf Auskunft gemäß Artikel 15 der Verordnung (EU) 2016/679 besteht darüber hinaus nicht, wenn die Daten für Zwecke der wissenschaftlichen Forschung erforderlich sind und die Auskunftserteilung einen unverhältnismäßigen Aufwand erfordern würde.

(3) Ergänzend zu den in § 22 Absatz 2 genannten Maßnahmen sind zu wissenschaftlichen oder historischen Forschungszwecken oder zu statistischen Zwecken verarbeitete besondere Kategorien personenbezogener Daten im Sinne des Artikels 9 Absatz 1 der Verordnung (EU) 2016/679 zu anonymisieren, sobald dies nach dem Forschungs- oder Statistikzweck möglich ist, es sei denn, berechtigte Interessen der betroffenen Person stehen dem entgegen. Bis dahin sind die Merkmale gesondert zu speichern, mit denen Einzelangaben über persönliche oder sachliche Verhältnisse einer bestimmten oder bestimmbaren Person zugeordnet wer-

In France, a draft of a new law on data protection (published in December 2017) does not include such exceptions for research in general (although a special framework concerns research in the domain of health). On the other hand, it implements exceptions for archiving in the public interest.[70]

### D. Appropriate safeguards?

As noted above, the precondition for application of the special rules concerning research is the application of "appropriate safeguards for the rights and freedoms of the data subject". These safeguards, as art. 89(1) of the GDPR further clarifies (in less-than-perfect language), "shall ensure that technical and organisational measures are in place in particular in order to ensure respect for the principle of data minimisation".

It seems that there is no fixed list of such safeguards; whether or not they are appropriate shall be evaluated on a case-by-case basis, taking into account in particular the character of the processed data (the more sensitive the data, the more advanced the safeguards) and the reasonable expectations of data subjects.

Such safeguards may include, but are not limited to[71]:

- pseudonymisation (see above), as expressly recognized by the GDPR;
- functional separation (i.e. taking measures to ensure that data are not used for other purpose than research, and in particular that the data are not used to take decisions or actions with respect to individuals);
- increased transparency (e.g. providing the data subjects with more information than actually required by law; also making the information publicly available);
- opt-out mechanisms, allowing data subjects to request removal of their data (even if their right to object is limited and the processing is not based on consent that could be withdrawn) or other mechanisms allowing the data subject to monitor the processing;

---

den können. Sie dürfen mit den Einzelangaben nur zusammengeführt werden, soweit der Forschungs- oder Statistikzweck dies erfordert.

(4) Der Verantwortliche darf personenbezogene Daten nur veröffentlichen, wenn die betroffene Person eingewilligt hat oder dies für die Darstellung von Forschungsergebnissen über Ereignisse der Zeitgeschichte unerlässlich ist.

It shall be noted that formally the Federal Data Protection Law (BDSG) only applies to university research to the extent that it is not regulated by a state data protection law, or LDSG (see art. 1(1) of the BDSG(neu)). It is therefore possible that the question will be regulated differently at the level of each state (Land).

[70] Art. 12 of the Projet de loi relatif à la protection des données personnelles (http://www.assemblee-nationale.fr/15/projets/pl0490.asp).

[71] Partially inspired by: Article 29 Data Protection Working Party, Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC, WP217, pp. 42-43.

- Data Protection Impact Assessment (see above), even if not required by law (to provide for reinforced accountability);
- the use of Privacy Enhancing Technologies (see above about data protection by design and by default);
- maintaining a detailed record of processing activities, exceeding what is required by law (reinforced accountability and transparency);
- the implementation of a robust, publicly available Data Management Plan and/or Privacy Policy;
- the approval of the processing of data for research purposes by an ethics committee (in cooperation with the Data Protection Officer or the competent data protection authority);
- encryption using state-of-the-art techniques;
- immediate deletion of data after use (reinforced principle of storage limitation);
- reinforced security, including e.g. access restrictions (only certain members of the research team can actually consult the personal data) or storage on a computer with no Internet connection;
- the adoption of an approved Code of Conduct or obtaining a Data Protection Seal (see below).

**E. Remark on "archiving in the public interest"**

The rules concerning "archiving in the public interest" are nearly exactly the same as those concerning research (see above B, C and D of this section).[72] It seems therefore that the same rules govern the use of personal data for research purposes and their long-term storage by specialised institutions, although some differences between Member States may persist due to the facultative nature of certain exceptions to rights of data subjects (see above about the national laws in Germany and in France).

It should be kept in mind, however, that personal data can be kept non-anonymised for research purposes (and therefore also archived) only when it is necessary to achieve the goals of the research.

[72] Art. 89(3) of the GDPR.

## III. New opportunities for bottom-up standardisation: Codes of Conduct and Data Protection Seals

The GDPR expressly recognizes the role of mechanisms such as codes of conduct, binding corporate rules or certificates, thereby creating interesting opportunities for bottom-up standardisation.

**Codes of Conduct**

Art. 40 of the GDPR encourages the adoption of Codes of Conduct "intended to contribute to the proper application of [the GDPR], taking account of the specific features of the various processing sectors". Such Codes of Conduct can be prepared by associations or other bodies representing categories of controllers (such as e.g. CLARIN ERIC) and specify various aspects related to data protection (e.g. the appropriate safeguards to be applied in research activities). The draft of such a Code of Conduct shall be submitted for approval to the competent supervisory authority. If the Code of Conduct relates to processing in several Member States (such as in the case of language research), the supervisory authority shall submit it further for approval by the European Data Protection Board; if approved, the Board transmits the Code of Conduct to the European Commission which may decide that the Code of Conduct has general validity in the European Union (i.e. even for data controllers who do not expressly adopt the Code) and therefore complement the GDPR.

Furthermore, approved Codes of Conduct may be monitored by a body which has "an appropriate level of expertise in relation to the subject-matter of the code and is accredited for that purpose by the competent supervisory authority".[73]

Such codes of conduct may also facilitate sharing of research data: if accepted by the recipient in a binding contract they may enable transfer of personal data to countries for which there are no adequacy decisions (see above).

The development of a GDPR Code of Conduct for language resources has been proposed by Kamocki et al. during the CLARIN Annual Conference of 2018.[74]

**Certificates (Data Protection Seals)**

Art. 42 of the GDPR encourages "the establishment of data protection certification mechanisms and of data protection seals and marks, for the purpose of demonstrating compliance with this Regulation". Such a certificate can be issued (for a maximum period of three years, renewable) by the competent supervisory authority or by an accredited certification body. The certificate does not reduce the responsibility of the controller or the processor for compliance with the GDPR and is without prejudice to the tasks and powers of the supervisory

[73] Art. 41 of the GDPR.

[74] Kamocki, Pawel et al., Toward a CLARIN Data Protection Code of Conduct, CLARIN Annual Conference 2018.

authorities. The European Data Protection Board shall keep a register of the existing certification mechanisms.

## Conclusion

The GDPR sets a high level of personal data protection by reinforcing and updating the principles already present in the Personal Data Directive and introducing certain new rules. It also allows some flexibility for research purposes, always provided that appropriate safeguards are applied. High administrative fines for non-respect of the GDPR and increased role of Data Protection Officers will undoubtedly draw more attention of the research community and of research funders to the issues of data protection.

Achieving compatibility with the GDPR will require considerable effort, especially in the preparatory phase, where such obligations as introducing data protection by design and by default or conducting Data Protection Impact Assessments need to be met. Anonymisation using appropriate, robust techniques may be seen as a remedy, since the GDPR does not apply to anonymous data; however, it would strip many types of language resources (containing audio or video) of most value for research. In such cases, alternative approaches to handling personal data while implementing appropriate safeguards need to be adopted.

The new framework will also need to be clarified via application in the years to come, as the text of the GDPR often is quite open to interpretation. It is important for the language research community to cooperate closely with Data Protection Officers and supervisory authorities to develop good practices.

Most importantly, the GDPR creates interesting opportunities for bottom-up standardisation: adoption of Codes of Conduct (which may be granted universal validity by the European Commission, thereby complementing the GDPR) or creation of certificates, marks and seals. It is important that the language research community makes the best of these opportunities.