Processing personal data without the consent of the data subject for the development and use of language resources

Aleksei Kelli
University of Tartu
Estonia
aleksei.kelli@ut.ee

Krister Lindén
University of Helsinki
Finland
krister.linden@helsinki.fi

Kadri Vider
University of Tartu
Estonia
kadri.vider@ut.ee

Pawel Kamocki
ELDA, France /
IDS Mannheim,
Germany
pawel.kamocki@gmail.com

Ramūnas Birštonas
Vilnius University
Lithuania
ramunas.birstonas@tf.vu.lt

Silvia Calamai
University of Siena
Italy
silvia.calamai@unisi.it

Chiara Kolletzek
Lawyer and Record Manager,
Italy
chiara.kolletzek@live.it

Penny Labropoulou
ILSP/ARC, Greece
penny@ilsp.gr

Maria Gavriilidou
ILSP/ARC, Greece
maria@ilsp.gr

Abstract

The development and use of language resources often involve the processing of personal data. The General Data Protection Regulation (GDPR) establishes an EU-wide framework for the processing of personal data for research purposes while at the same time it allows for some flexibility on the part of the Member States. The paper discusses the legal framework for language research following the entry into force of the GDPR. To this goal, we first present some fundamental concepts of data protection relevant for language research and then focus on the models that certain EU member states use to regulate data processing for research purposes.

1 Introduction

Language resources contain material subject to various legal regimes. For instance, language resources can contain copyright protected works, objects of related rights (performances) and personal data. This affects the way language resources are collected and used. Intellectual property issues relating to language resources have been previously addressed (see Kelli et al. 2015). The focus of this article is on personal data protection. More precisely on the processing of personal data for research purposes without the consent of the data subject within the framework of language research. Personal data issues are relevant for language resources, given that they potentially contain oral speech or written text which relates to a natural person.[1] In the CLARIN Virtual Language Observatory (VLO), 95,502 language resources[2] could contain personal data.[3]

Aleksei Kelli, Krister Lindén, Kadri Vider, Pawel Kamocki, Ramūnas Birštonas, Silvia Calamai, Chiara Kolletzek, Penny Labropoulou and Maria Gavrilidou 2019. Processing personal data without the consent of the data subject for the development and use of language resources. *Selected papers from the CLARIN Annual Conference 2018*. Linköping Electronic Conference Proceedings 159: 72–77.

Although the General Data Protection Regulation[4] (GDPR) provides a general framework for personal data protection, it leaves a certain degree of freedom for the EU Member States to regulate data processing for research purposes. It means that they can adopt different regulatory models. This article preliminarily maps and provides insights into different models.[5] Before concentrating on the data processing for research purposes, key concepts of the data protection framework are addressed.

2 Data subject, personal data and data processing

The data subject is defined through the concept of personal data. Personal data is "any information relating to an identified or identifiable natural person ('data subject')" (GDPR Art. 4). Publicly available personal data is also protectable (C-73/07). According to Article 29 Working Party[6] (WP29), information contained in free text in an electronic document may qualify as personal data. It does not have to be in a structured database (2007: 8).

The identifiability is a crucial issue since data not relating to a natural person (incl. anonymous data) is not subject to the GDPR requirements (See GDPR Recital 26). A natural person can be identified by reference to the identifier (name, identification number), location data and physiological, genetic, mental, economic, cultural or social information (GDPR Art. 4). According to WP29 sound and image data qualify as personal data insofar as they may represent information on an individual (WP29 2007: 7). It means that LRs containing oral speech are subject to the GDPR. A question can be raised whether speech and voice as such constitute personal data where there is no additional information leading to a specific individual. It is a question related to identifiability. As suggested in the literature, data that are not identifiable for one person may be identifiable for another. Data can also become identifiable through combination with other data sets. Identifiability is a broad category depending on how much effort must be deemed 'reasonable' (Oostveen 2016: 306).

Voice can be considered biometric data (see González-Rodríguez et al. 2008; Jain et al. 2004).[7] Biometric data for uniquely identifying a natural person belongs to a special category of personal data[8] the processing of which is even more restricted than for other personal data. The similar case is with photos depicting people. Here the GDPR provides a clarification: "The processing of photographs should not systematically be considered to be processing of special categories of personal data as they are covered by the definition of biometric data only when processed through a specific technical means allowing the unique identification or authentication of a natural person" (Recital 51). This should be applicable in case of speech as well. Therefore, the requirements concerning the processing of special categories of personal data do not need to be followed until the oral speech contained in language resources is not used for the identification of natural persons.

The GDPR defines processing very broadly. It includes, among other things, collection, structuring, storage, adaptation, use, making available or destruction (GDPR Art. 4). It means that the development and use of LRs containing personal data constitutes processing.

---

[1] For instance, according to the Court of Justice of the European Union (CJEU) the concept of personal data covers the name of a person (C-101/01).

[2] Resource type: Audio, Radio, Sound, Speech, Spontaneous, Television or Video.

[3] Language resources with written text may also contain personal data, but this is not as prominent as in the case of audio and/or visual material (e.g. interviews or photos of a certain person).

[4] The GDPR is applicable in all EU Member States from 25 May 2018. It replaces the Data Protection Directive.

[5] For lack of space not all the EU countries are addressed in the present paper.

[6] According to the Data Protection Directive the Working Party on the Protection of Individuals with regard to the Processing of Personal Data (WP29) is composed of a representative of the supervisory authority or authorities designated by each Member State and of a representative of the authority or authorities established for the Community institutions and bodies, and of a representative of the Commission.

[7] The GDPR defines biometric data as "personal data resulting from specific technical processing relating to the physical, physiological or behavioural characteristics of a natural person, which allow or confirm the unique identification of that natural person, such as facial images or dactyloscopic data" (Art. 4).

[8] The GDPR defines special categories of personal data as "data revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation".

3 Processing personal data for research purposes

3.1 GDPR requirements

The GDPR provides six legal grounds for processing personal data: consent, performance of a contract, compliance with a legal obligation, protection of the vital interests, public interest and legitimate interests (Art. 6). If possible, the processing for research purposes should rely on consent (for further discussion on consent see WP29 2017). This paper concentrates on cases where there is no consent, and another legal ground is needed.[9] According to WP29, the legitimate interest can serve as a legal ground for processing personal data in the research context (2014b: 24-25). The concept of legitimate interests is rather complicated and requires weighing different interests.[10]

The GDPR establishes the following requirements for processing[11] data in the research[12] context (Art. 89):

1) processing is subject to appropriate safeguards (technical and organisational measures (e.g., pseudonymisation) to ensure data minimisation);

2) the EU Member States are allowed to limit the following data subject's rights[13]: the right of access, right to rectification, right to restriction of processing and right to object.

Since the GDPR offers the Member States some flexibility to specify requirements concerning processing for research purposes[14], it is necessary to analyse national laws.

3.2 National models

The first example to be considered is **Estonia**. The Estonian draft Act on Personal Data Protection (Draft PDPA 2018a) sets the following requirements for processing of personal data for scientific research (§ 6):

1) Personal data may be processed without the consent of the data subject for research purposes mainly if data has undergone pseudonymisation.

2) Processing of data without consent for scientific research in a format which enables identification of the data subject is permitted only if the following conditions are met:

a) after removal of the data enabling identification, the goals of data processing would not be achievable, or achievement thereof would be unreasonably difficult;

b) the person carrying out the scientific research finds that there is a predominant public interest for such processing;

c) obligations of the data subject are not changed by the processed personal data, and the rights of the data subject are not excessively damaged in any other manner.

3) The data controller may limit the data subject's right of access, right to rectification, right to restriction of processing and right to object in so far as the exercise of these rights are likely to render impossible or seriously impair the achievement of the objectives of the processing for research purposes.

4) In case of processing of special categories of personal data an ethics committee in the corresponding area verifies, before the commencement of the processing, compliance with the requirements set out in this section. In the absence of an ethics committee in a specific area, the Data Protection Authority verifies the fulfilment of requirements.

[9] As a matter of fact, one option to avoid problems with personal data protection is the anonymisation of data used for language research (for further discussion on anonymization of data see WP29 2014a).

[10] According to the GDPR 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 which require protection of personal data" (Art. 6).

[11] Including further processing of previously collected data (Art. 5).

[12] The GDPR defines research in a broad manner. It covers "technological development and demonstration, fundamental research, applied research and privately funded research" (Recital 159).

[13] The GDPR itself limits the right to erasure ('right to be forgotten') in research context (Art. 17).

[14] MS have flexibility to regulate data processing in other areas as well (e.g., processing for the purpose of academic artistic or literary expression (Art. 85)).

According to the **Finnish** model, the draft Act on Personal Data Protection (Draft PDPA 2018b) and its preamble outline the following for processing personal data for scientific research[15]:

1) Personal data may be processed by university researchers according to § 6.1e in the GDPR, i.e. *performance of a task carried out in the public interest* based on the university's legal mandate to do research. Universities also have the right to archive data for scientific and historical research based on §9.2j in GDPR, i.e. *processing is necessary for archiving purposes in the public interest, scientific or historical research purposes or statistical purposes*.

2) For a researcher to use datasets containing personal data for secondary purposes, a research plan is required, but its primary purpose from the GDPR perspective is to document that the research is of a scientific or historical nature. If the personal data is sensitive, a set of protective measures specified in the law also need to be carried out and documented.

3) Also, the data subjects may have limited rights to stop processing of personal data for scientific and historical research if the processing is necessary for carrying out the research. The motivation for why the processing is necessary should be included in the research plan (mentioning the Principal Investigator). It is possible even for sensitive personal data, but then the research plan also needs to describe how it fulfils the ethical standards in the field of research and needs to be delivered to the Data Protection Authority at least 30 days ahead of starting the processing.

The next example is **Lithuania**. In contrast with the current regulation, the newly enacted Lithuanian Law Amending the Law on Legal Protection of Personal Data (LLPPD 2018), which implements the provisions of GDPR, contains no special provisions dealing with the research exemption and no prior checking procedure is required. It means that Lithuania has not used the opportunities and flexibilities provided in Art. 89 of GDPR. It also means that after the implementation of GDPR, the persons using personal data for scientific research has to rely directly on and comply with the general provisions of GDPR.

Italy has not yet released detailed legislation to link the Italian Personal Data Protection Code (IPDPC) to the GDPR. However, the subject is addressed at the policy level. The Law of European Delegation (LED) establishes that the Italian Government enacts detailed legislative decrees with the aim of adapting the national framework specifying the GDPR (art. 13). In the Delegation Law, a series of objectives are defined: (i) the abrogation of the provisions of the Privacy Code which are not compatible with the provisions of the GDPR; (ii) the coordination and integration of the Code on personal data protection, in order to implement the non-directly applicable GDPR provisions; (iii) the adoption of specific implementing Acts by the Italian data protection Authority (i.e. "Garante Privacy"), for the purposes envisaged by the GDPR.

It is likely that clause (iii) also deals with the use of data for research purposes. The data protection authority assesses whether the provisions contained in Attachment A.2 of the Privacy Code are fully compliant to the GDPR rules, or whether the additional regulatory action is necessary.

France and **Germany** have adopted very different views on the processing for research purposes. The new German law[16] seems to be as favourable to researchers as possible, while the new French law is much more conservative. For example, the French law requires authorisation from the national data protection authority for processing of special categories of personal data for public research purposes, whereas the German law generally allows for such data to be processed for such purposes simply if the processing is subject to appropriate safeguards and if it passes the "balance of interests" test. Likewise, any derogations from rights of data subjects in the French law seem to be limited to specific cases of medical research. Time shall tell which approach proves better.

[15] As far as possible, Finland plans to maintain its existing practice for collecting and using research data.

[16] It shall be kept in mind that the German Federal Data Protection Act (Bundesdatenschutzgesetz, BDSG) only applies to processing of personal data by private entities and by public bodies of the German Federation (art. 1 of the BDSG). Processing of personal data by public bodies of the Länder (such as universities) is governed by regional norms (Landesdateschutzgesetze, LDSG). To the best of our knowledge, no LDSGs has yet been updated to conform to the GDPR. Therefore, for now the situation regarding processing of personal data for research purposes in German universities is not entirely clear.

In **Greece**, a Draft Bill for Personal Data (PDPA 2018c) implementing the GDPR has been recently released for public consultation (completed on March 5, 2018). The Bill contains an article dedicated to the processing of PD for "scientific or historical research or for statistical data". Processing of PD is allowed *if the subjects have given their consent for this or previous studies on the same data, if the data come from publicly accessible sources* or *if the processing can be proven to be required for the research*. For the processing of the special categories, the Bill is more restrictive; especially for research on genetic data prior consultation with the Data Protection Authority is mandatory. Medical data processing is allowed, provided the researchers involved are legally or professionally bound by confidentiality. Pseudonymisation or anonymisation are recommended but only when they do not hinder the purposes of the research. Overall, this draft Bill can be considered favourable towards research purposes.

4 Conclusion

As argued in the paper, the development and use of language resources often involve processing of personal data. Although the GDPR is applicable in the whole EU, it allows the Member States to specify processing for research purposes. This means that in addition to the GDPR, researchers that wish to construct and use LRs for language research must further follow national requirements.

Reference

[BDSG] Bundesdatenschutzgesetz. Available at https://www.gesetze-im-internet.de/bdsg_2018/index.html (5.9.2018)

[C-101/01] Case C-101/01. Criminal proceedings against Bodil Lindqvist (6 November 2003). Available at http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1521039149443&uri=CELEX:62001CJ0101 (3.4.2018)

[C-73/07] Case C-73/07. Tietosuojavaltuutettu vs. Satakunnan Markkinapörssi Oy and Satamedia Oy (16 December 2008). Available at https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:62007CA0073&qid=1536154290371&from=EN (5.9.2018)

[Data Protection Directive] Directive 95/46/EC of the European Parliament and of the Council of 24 October 1995 on the protection of individuals with regard to the processing of personal data and on the free movement of such data. Official Journal L 281, 23/11/1995 p. 0031 – 0050. Available at http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:31995L0046&qid=1522340616101&from=EN (29.3.2018)

[Draft PDPA 2018a] Estonian Draft Act on Personal Data Protection (Isikuandmete kaitse seaduse eelnõu) (22.08.2018). Available at https://www.riigikogu.ee/tegevus/eelnoud/eelnou/5c9f8086-b465-4067-841e-41e7df3b95af/Isikuandmete%20kaitse%20seadus (3.4.2018)

[Draft PDPA 2018b] Finnish Draft Act on Personal Data Protection (Hallituksen esitys eduskunnalle EU:n yleistä tietosuoja-asetusta täydentäväksi lainsäädännöksi) (01.03.2018). Available at https://www.eduskunta.fi/FI/vaski/HallituksenEsitys/Sivut/HE_9+2018.aspx (4.4.2018)

[Draft PDPA 2018c] Greek Draft Bill on Personal Data Protection (Νόμος για την Προστασία Δεδομένων Προσωπικού Χαρακτήρα). Available at http://www.opengov.gr/ministryofjustice/wp-content/uploads/downloads/2018/02/sxedio_nomou_prostasia_pd.pdf (18.4.2018)

[French law] Loi n° 2018-493 du 20 juin 2018 relative à la protection des données personnelles, modifying the French Data Protection Act (loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés)

[GDPR] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). OJ L 119, 4.5.2016, p. 1-88. Available at http://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1515793631105&uri=CELEX:32016R0679 (29.3.2018)

[González-Rodríguez et. al. 2008] Joaquín González-Rodríguez, Doroteo Torre Toledano, Javier Ortega-García (2008). Voice Biometrics. In Handbook of Biometrics edited by Anil K. Jain, Patrick Flynn, Arun A. Ross. Springer

[IPDPC] Italian Personal Data Protection Code. Legislative Decree 30.06.2003 No. 196. English version available at: http://194.242.234.211/documents/10160/2012405/Personal+Data+Protection+Code+-+Legislat.+Decree+no.196+of+30+June+2003.pdf (11.4.2018)

[Jain et. al. 2004] Anil K. Jain, Arun Ross, Salil Prabhakar (2004). An Introduction to Biometric Recognition. - IEEE Transactions on Circuits and Systems for Video Technology 14(1). Available at https://www.cse.msu.edu/~rossarun/BiometricsTextBook/Papers/Introduction/JainRossPrabhakar_BiometricIntro_CSVT04.pdf (31.3.2018)

[Kelli et al. 2015] Aleksei Kelli, Kadri Vider, Krister Lindén (2015). The Regulatory and Contractual Framework as an Integral Part of the CLARIN Infrastructure. 123: Selected Papers from the CLARIN Annual Conference 2015, October 14–16, 2015, Wroclaw, Poland. Ed. Koenraad De Smedt. Linköping University Electronic Press, Linköpings universitet, 13−24. Available at http://www.ep.liu.se/ecp/article.asp?issue=123&article=002 (28.3.2018)

[LED] Law of European Delegation. Law No. 25.10.2017 No 163. Available at http://www.gazzettaufficiale.it/eli/id/2017/11/6/17G00177/sg (11.4.2018)

[LLPPD 2018] Lithuanian Law Amending the Law on Legal Protection of Personal Data (Lietuvos Respublikos asmens duomenų teisinės apsaugos įstatymo pakeitimo įstatymas). Available at https://www.e-tar.lt/portal/legalAct.html?documentId=43cddd8084cc11e8ae2bfd1913d66d57 (30.8.2018)

[Oostveen 2016] Manon Oostveen (2016). Identifiability and the applicability of data protection to big data. International Data Privacy Law 6 (4), 299-309

[Personal Data Protection Act]. Personal Data Protection Act. Entry into force 01.01.2008. English translation available at https://www.riigiteataja.ee/en/eli/507032016001/consolide (3.4.2018)

[Privacy Code] Code of conduct and professional practice Regarding the processing of personal data for historical purposes. English version available at http://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/export/1565819 (11.4.2018)

[VLO] CLARIN Virtual Language Observatory. Available at https://vlo.clarin.eu/ (18.4.2018)

[WP29 2017] WP29. Guidelines on Consent under Regulation 2016/679. Adopted on 28 November 2017 [adopted, but still to be finalized]. Available at http://ec.europa.eu/newsroom/article29/item-detail.cfm?item_id=615239 (2.4.2018)

[WP29 2014a] WP29. Opinion 05/2014 on Anonymisation Techniques Adopted on 10 April 2014. Available at http://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp216_en.pdf (3.4.2018)

[WP29 2014b] WP29. Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC. Available at http://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp217_en.pdf (3.4.2018)

[WP29 2007] WP29. Opinion 4/2007 on the concept of personal data. Adopted on 20th June. Available at http://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf (29.3.2018)