The Interplay of Legal Regimes of Personal Data, Intellectual Property and Freedom of Expression in Language Research

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

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

Pawel Kamocki
IDS Mannheim, Germany
kamocki@ids-mannheim.de

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

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

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

Vadim Mantrov
University of Latvia, Latvia
vadims.mantrovs@lu.lv

Vanessa Hannesschläger
OeAW, Austria
vanessa.hannesschlaeger gmail.com

Riccardo Del Gratta
ILC, Italy
riccardo.delgratta ilc.cnr.it

Age Värv
University of Tartu, Estonia
age.varv@ut.ee

Gaabriel Tavits
University of Tartu, Estonia
gaabriel.tavits@ut.ee

Andres Vutt
University of Tartu, Estonia
andres.vutt@ut.ee

Abstract

Sometimes legal scholars get relevant but baffling questions from laypersons like: "*The reference to a work is personal data, so does the GDPR actually require me to anonymise it? Or, as my voice data is personal data, does the GDPR automatically give me access to a speech recognizer using my voice sample? Or, can I say anything about myself without the GDPR requiring the web host to anonymise or remove the post? What can I say about others like politicians? And, what can researchers say about patients in a research report?*" Based on these questions, the authors address the interaction of intellectual property and data protection law in the context of data minimisation and attribution rights, access rights, trade secret protection, and freedom of expression.

1 Introduction

There is an awareness that intellectual property (IP) and personal data (PD) protection are relevant in language research. These two regimes are often applicable simultaneously, and their requirements might

This work is licenced under a Creative Commons Attribution 4.0 International Licence. Licence details: http:// creativecommons.org/licenses/by/4.0/

seem contradictory. Therefore, the authors have chosen three specific cases[1] to outline the interaction of IP and PD protection and provide preliminary guidance.

Firstly, the authors explore the interplay between the data minimisation principle and the right to be acknowledged as the author (the attribution right). On the one hand, the data minimisation principle enshrined in the General Data Protection Regulation (GDPR) requires processing[2] as little PD as possible (Art. 5 (1) c)). On the other hand, the Berne Convention Art. 6bis gives the authors the attribution right. The relevant question here is whether a researcher who has collected language data (LD) containing copyrighted content should attribute the author of the content or follow the data minimisation principle and remove all PD (e.g., the author's name) that is not necessary for processing.

The second case concerns IP protection and the data subject's access right. In practical terms, a researcher might need to decide what data the access right covers. Is it only raw PD or data derived from PD (e.g., a language model which could contain trade secrets)?

Thirdly, the authors discuss the impact of PD protection on freedom of expression since publications constitute research outcomes. The authors rely on the regulation of their countries.

2 The data minimisation and the right of attribution

According to the data minimisation principle, PD must be "*adequate, relevant and limited to what is necessary in relation to the purposes for which they are processed*" (GDPR Art. 5 (1) clause c).

The European Data Protection Board (EDPB) explains it further: "*minimising can also refer to the degree of identification. If the purpose of the processing does not require the final set of data to refer to an identified or identifiable individual (such as in statistics), but the initial processing does (e.g., before data aggregation), then the controller shall anonymize personal data as soon as identification is no longer needed. Or, if continued identification is needed for other processing activities, personal data should be pseudonymized to mitigate risks for the data subjects' rights*" (2019: 19).

According to Art. 6bis (1) of the Berne Convention, "*the author shall have the right to claim authorship of the work*". The InfoSoc Directive also contains the obligation to identify the source (incl. the author's name) (Art. 5 (3)). The EU case law reiterates the obligation (e.g., C-145/10). In other words, there is a legal obligation to acknowledge the author of the content. It is compatible with the GDPR since it names compliance with a legal obligation as a legal basis for PD processing (Art. 6 (1) c)).

An overarching theme for this and the following section concerns legal obligations relating to derived data (e.g., data derived through text and data mining (TDM). For further discussion, see Kelli et al. (2020). Interestingly, the TDM exception contained in the DSM Directive does not require attribution. However, it should be borne in mind that the TDM exception only limits the reproduction right. In case the results of TDM are disseminated, the attribution right has to be honoured.[3]

3 The scope of access right and trade secret protection

In combination with the right to be informed and the principle of transparency, the access right forms a foundation for exercising the data subjects' rights. The access right requires the controller to provide information on the processing of PD (GDPR Art. 15). The first question for research organisations and researchers (data controllers) is the scope of PD subject to the access right. The data subjects should have access to their raw data. The question is whether the access right applies to data derived from PD as well. This question asks what PD covers.[4] The Court of Justice of the European Union (CJEU) has not been remarkably consistent. For instance, it has explained that "*There is no doubt that the data relating to the applicant for a residence permit and contained in a minute, such as the applicant's name, date of birth, nationality, gender, ethnicity, religion and language, are […] 'personal data' […] As*

[1] It should be mentioned that there are a myriad of IP and PD protection interaction points whose systematic mapping is outside the scope of this article. Therefore, the authors chose cases which could potentially be relevant for language researchers.

[2] The GDPR defines processing so widely that it covers all possible activities with PD (Art. 4 (2)).

[3] The attribution right exists only in case of the existence of copyrighted content. In the EU case law, it is pointed out that 11 consecutive words could be copyright protected (C-5/08). However, it does not say that less than 11 words are not copyrighted. For further discussion, see Kamocki 2020.

[4] For the concept of PD, see WP29 2007.

regards, on the other hand, the legal analysis in a minute, it must be stated that, although it may contain personal data, it does not in itself constitute such data" (C-141/12 paragraphs 38, 39). In another case, the CJEU held that "*the written answers submitted by a candidate at a professional examination and any comments made by an examiner with respect to those answers constitute personal data*" (C-434/16).

Understandably, the concept of PD should be interpreted extensively. However, there is no legal clarity on whether data derived from PD should be made available as well. WP29 (2016: 9) suggests in the context of the right of portability (see GDPR Art. 20) that "*user categorisation or profiling are data which are derived or inferred from the personal data provided by the data subject, and are not covered by the right to data portability*".

Within the context of language research, the question is whether the data subject could require access to a language model which was trained using his PD. If not, there should be a legal basis limiting the scope of access right. One argument here is that model is protected by intellectual property (copyright, trade secret). The GDPR accepts this line of argument in its Recital 63, explaining the nature of the access right: "*That right should not adversely affect the rights or freedoms of others, including trade secrets or intellectual property and in particular the copyright protecting the software. However, the result of those considerations should not be a refusal to provide all information to the data subject*".

To sum up, the access right does not cover the access to language models, especially when their creators consider trade secrets. In this context, IP rights prevail over data protection.

4 Data subject's rights and freedom of expression

The data subject has the right to object to the processing and obtain the erasure, restriction or rectification of PD concerning him (GDPR Art. 16, 17, 18, 21). These rights may conflict with the author's right to make his work available. This question can be framed as an interaction of PD protection and freedom of expression (FoE). Protection of PD is not an absolute right (GDPR Rec. 4). Therefore, the GDPR allows Member States to limit the data subject's rights to reconcile PD protection with the freedom of expression.

The EU countries have followed different implementation routes. For instance, the German Federal Data Protection Act (BDSG) does not contain any rules specifically implementing Article 85 of the GDPR. The existing broad derogation for research and archiving purposes (Article 27 of the BDSG) is based on Article 89, not 85 of the GDPR. It seems to be deemed sufficient by the legislator (Deutscher Bundestag, 2018). Specific state acts regarding media and journalistic expression exist in many federal states (Länder), e.g., Hessisches Pressegesetz or Landesmediengesetz Baden-Württemberg.

Most of the studied countries have adopted a general provision limiting the applicability of the GDPR to ensure freedom of expression (see Austrian, Estonian, Finnish and Lithuanian PDPA). France, Greece, Italy, and Latvia have more detailed regulations, which are explored below.

Article 80 of the French Data Protection Act attempts to reconcile data protection and freedom of expression in France. It derogates from two general principles: the storage limitation and the prohibition of processing sensitive data (including data about criminal convictions and offences). It also limits information rights, access, rectification and restriction, and derogates from the rules on data transfers. This framework applies only when necessary to safeguard freedom of expression and information, and only when the data are processed: 1) for academic (*'universitaire'*), artistic or literary expression, or 2) for journalistic purposes by professional journalists, in a way that respects ethical rules (deontology) of the profession. The Article clearly states that other laws and codes regarding violations of privacy and reputational damage continue to apply.

One can be surprised by the adjective '*universitaire'* in Article 80 of the French Copyright Act (expression *universitaire, artistique ou littéraire*) rather than '*académique'* (as in '*academic, artistic or literary expression*'). However, the same wording is used by the French version of Article 85 of the GDPR. This is because '*académique'* has a very restricted meaning in French (related to the Académie Française) and should not be interpreted as limiting the derogatory framework to processing made by scholars with a university affiliation.

Article 28 of the Greek Personal Data Protection Act, corresponding directly to the GDPR (Art. 85), aims to reconcile the right to personal data protection with the right to freedom of expression and information, *"including the processing for journalistic purposes and for purposes of academic, literary or artistic expression".* More specifically, in the framework of these objectives, Paragraph 1 of this Article

explicitly enumerates cases where the processing of PD is allowed: *"(a) when the subject of the data has given his explicit consent, (b) for PD that have been publicised by the subject, (c) when the right to the freedom of expression and the right to information overweighs the right to PD protection, especially for topics of general interest or when the PD relates to public persons, and (d) when it is restricted to the necessary measure to ensure the right of expression and the right of information, especially with regard to sensitive categories of PD, and criminal cases, and security related measures, taking into account the right of the subject to his private and family life."* We can deduce that the Article looks more into the '*journalistic purposes*' rather than '*academic purposes*'. Paragraph 2 of the same Article provides the exceptions and derogations for processing for such purposes, which are mentioned in Article 85 of the GDPR.

Article 136 of the Italian Personal Data Protection Code (PDPC) implements Art. 85 of the GDPR. It regulates journalistic as well as academic works. Article 137 defines the categories of PD that can be processed without the data subject's consent. Namely, such categories are special categories of PD and PD data related to criminal convictions and offences (GDPR Art. 9, 10). Other sections further restrict these categories to *"Safeguards applying to the processing of genetic data, biometric data, and data relating to health"* (section 2-f) and *"Processing entailing a high risk for the performance of a task carried out in the public interest"* (section 2-p). Article 137 (3) provides: "*It shall be allowed to process the data concerning circumstances or events that have been made known (communicated/disseminated) either directly by the data subject or on account of the data subject's public conduct*".

Article 32(3) of the Latvian PDPA states that when processing data for academic, artistic or literary expression, provisions of the GDPR (except for Article 5) shall not be applied if all of the following conditions are present: 1) data processing is conducted by respecting the right of a person to private life, and it does not affect interests of a data subject which require protection and override the public interest; 2) compliance with the provisions of the GDPR is incompatible with or prevents the exercise of the rights to freedom of expression and information.

There are two intriguing questions concerning the interaction of PD protection and freedom of expression:

1) how to strike a fair balance between PD protection and FoE in research settings? FoE is usually framed in the context of newspapers publishing facts about public figures. The freedom of academic expression is somewhat unclear. Still, it has been interpreted to apply to, e.g., x-ray pictures of medical case studies as standard practice. Such accompanying material is publicly disclosed in a scientific journal as an illustration of the published case. There is no need to obtain the consent of the x-rayed person for this purpose. Usually, a person cannot be directly identified from such an x-ray. However, if the medical condition is rare, the individual may still be identifiable with the help of additional information. As the GDPR defines data concerning health as special categories of PD (Art. 9), this is an especially delicate example.

2) how and where to draw a line between processing for academic expression and research purposes. Research publication requires prior research. The question is whether this research is covered with the freedom of academic expression. The authors admit that when data is present in the research publication, then the processing could be covered by the FoE exception. It is important to emphasise that the principles of data minimisation, purpose limitation, accuracy, fairness (GDPR Art. 5), and other requirements need to be followed. At the same time, research ethics and funders' requirements require the publication of research data to ensure research reproducibility and verifiability. Therefore, there is a tension between open data and personal data protection requirements. For further discussion, see Kelli et al. (2018).

PD protection and FoE are both human rights. This means that one is not prioritized over another. The key issue is to strike a fair balance between them. PD protection should not affect academic freedom of expression.

5 Conclusion

The authors' reached the following preliminary conclusions. Firstly, the data minimisation and the attribution right are not contradictory concepts. The acknowledgement of the author is compatible with the GDPR as the compliance with a legal obligation. The attribution does not concern all PD but only data that is copyrighted. Secondly, the access right primarily applies to raw PD. There is no legal clarity regarding the access to data derived from PD. The access right does not presumably cover language

models containing trade. Thirdly, PD protection usually does not take precedence over the freedom of expression and cannot hinder the academic FoE and the author's right to disseminate his work. Conducting research could be covered FoE.

References

Austrian PDPA. *Data Protection Amendment Act.* Entry into force 2018. Available at https://www.ris.bka.gv.at/eli/bgbl/I/2018/31 (3.4.2021).

Berne Convention. *Berne Convention for the Protection of Literary and Artistic Works of September 9, 1886*. Available at https://wipolex.wipo.int/en/text/283698 (3.4.2021).

C-434/16. *Case C-434/16.* Peter Nowak v Data Protection Commissioner (20 December 2017). Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62016CJ0434&qid=1617649690306 (5.4.2021).

C-141/12. *Joined Cases C-141/12 and C-372/12.* YS (C-141/12) vs Minister voor Immigratie, Integratie en Asiel, and Minister voor Immigratie, Integratie en Asiel (C-372/12) v M, S (17 July 2014). Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62012CJ0141&qid=1617633666683 (5.4.2021).

C-145/10. *Case C-145/10.* Eva-Maria Painer vs Standard VerlagsGmbH and Others (1 December 2011). Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A62010CJ0145&qid=1618044850444 (10.4.2021).

C-5/08. *Case C-5/08.* Infopaq International A/S vs Danske Dagblades Forening (16 July 2009). Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1555243488182&uri=CELEX:62008CJ0005 (10.4.2021).

Deutscher Bundestag, 2018. *Deutscher Bundestag,* Ausarbeitung: Die Öffnungsklausel des Art. 85 der Datenschutz-Grundverordnung, WD 3 - 3000 - 123/18. Available at https://www.bundestag.de/resource/blob/560944/956f5930221c807984d40c1df2af5abf/WD-3-123-18-pdf-data.pdf (26.04.2021).

DSM Directive. Directive (EU) 2019/790 of the European Parliament and of the Council of 17 April 2019 on copyright and related rights in the Digital Single Market and amending Directives 96/9/EC and 2001/29/EC. *OJ L 130, 17.5.2019*, pp. 92-125. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1572352552633&uri=CELEX:32019L0790 (10.4.2021).

EDPB 2019. *European Data Protection Board.* Guidelines 4/2019 on Article 25 Data Protection by Design and by Default. Adopted on 13 November 2019. Available at https://edpb.europa.eu/sites/edpb/files/consultation/edpb_guidelines_201904_dataprotection_by_design_and_by_default.pdf (4.4.2021).

Estonian Copyright Act. *Copyright Act.* Entry into force 12.12.1992. Available at https://www.riigiteataja.ee/en/eli/504032021006/consolide (3.4.2021).

Estonian PDPA. *Personal Data Protection Act.* In force from: 15.01.2019. Available at https://www.riigiteataja.ee/en/eli/523012019001/consolide (6.4.2021).

EU Charter of Fundamental Rights. *Charter of Fundamental Rights of the European Union.* 2012/C 326/02. OJ C 326, 26.10.2012, p. 391-407. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:12012P/TXT (1.4.2021).

Finnish PDPA. *Data Protection Act (1050/2018).* Available at https://www.finlex.fi/en/laki/kaannokset/2018/en20181050.pdf (8.4.2021).

French Data Protection Act. *Loi n° 78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés*. Available at https://www.legifrance.gouv.fr/loda/id/JORFTEXT000000886460/ (27.4.2021).

French Intellectual Property Code. *Code de la propriété intellectuelle.* Available at https://www.legifrance.gouv.fr/codes/texte_lc/LEGITEXT000006069414/ (27.4.2021).

German Federal Data Protection Act. *Bundesdatenschutzgesetz vom 30. Juni 2017 (BGBl. I S. 2097)*, das durch Artikel 12 des Gesetzes vom 20. November 2019 (BGBl. I S. 1626) geändert worden ist. Available at: https://www.gesetze-im-internet.de/bdsg_2018/BJNR209710017.html (27.4.2021).

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 https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1555312258399&uri=CELEX:32016R0679 (1.4.2021).

Greek Data Protection Act. *Personal Data Protection Law (4624/2019)*. Available at: https://www.e-nomothesia.gr/kat-dedomena-prosopikou-kharaktera/nomos-4624-2019-phek-137a-29-8-2019.html (26.4.2021).

InfoSoc Directive. Directive 2001/29/EC of the European Parliament and of the Council of 22 May 2001 on the harmonisation of certain aspects of copyright and related rights in the information society. *Official Journal L 167*, 22/06/2001 P. 0010 – 0019. Available at https://eur-lex.europa.eu/legal-content/EN/TXT/?qid=1555254956114&uri=CELEX:32001L0029 (4.4.2021).

Italian PDPA. *Personal Data Protection Code containing provisions to adapt the national legislation to 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. Available at https://www.garanteprivacy.it/documents/10160/0/Data+Protection+Code.pdf/7f4dc718-98e4-1af5-fb44-16a313f4e70f?version=1.3 (8.4.2021).

Kamocki, Pawel. 2020. When Size Matters. Legal Perspective(s) on N-grams. *Proceedings of CLARIN Annual Conference 2020. 05 – 07 October 2020*. Virtual Edition. Ed. Costanza Navarretta, Maria Eskevich. CLARIN, 166-169. Available at https://office.clarin.eu/v/CE-2020-1738-CLARIN2020_ConferenceProceedings.pdf (10.4.2021).

Kelli, Aleksei, Arvi Tavast, Krister Lindén, Kadri Vider, Ramunas Birštonas, Penny Labropoulou, Irene Kull, Gaabriel Tavits, Age Värv, Pavel Stranák, Jan Hajic. 2020. The Impact of Copyright and Personal Data Laws on the Creation and Use of Models for Language Technologies. In: Kiril Simov, Maria Eskevich (Ed.). *Selected Papers from the CLARIN Annual Conference 2019 (53−65)*. Linköping University Electronic Press. Available at https://ep.liu.se/ecp/172/008/ecp20172008.pdf (10.4.2021).

Kelli, Aleksei, Tõnis Mets, Lars Jonsson, Krister Lindén, Kadri Vider, Ramūnas Birštonas, Age Värv. 2018. Challenges of Transformation of Research Data into Open Data: the Perspective of Social Sciences and Humanities. *International Journal of Technology Management and Sustainable Development*, 17 (3), 227−251.

Latvian Copyright Act. *Latvian Copyright Act.* Available at https://vvc.gov.lv/image/catalog/dokumenti/Copyright%20Law.doc (26.4.2021).

Latvian PDPA. *Latvian Personal Data Processing Law.* Available at https://vvc.gov.lv/image/catalog/dokumenti/Personal%20Data%20Processing%20Law.doc (26.4.2021).

Lithuania PDPA. *Republic of Lithuania Law on Legal Protection of Personal Data.* Available at https://www.e-tar.lt/portal/legalAct.html?documentId=43cddd8084cc11e8ae2bfd1913d66d57 (26.4.2021).

WP29 2016. *Article 29 Working Party (WP29)*. Guidelines on the right to data portability. Adopted on 13 December 2016. Available at https://ec.europa.eu/information_society/newsroom/image/document/2016-51/wp242_en_40852.pdf (5.4.2021).

WP29 2007. *Article 29 Working Party (WP29)*. Opinion 4/2007 on the concept of personal data. Adopted on 20th June. Available at https://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2007/wp136_en.pdf (5.4.2021).