Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 3423–3427
Marseille, 11–16 May 2020

Privacy by Design and Language Resources

Pawel Kamocki, Andreas Witt
Leibniz Institut für Deutsche Sprache, Leibniz Institut für Deutsche Sprache/CLARIN ERIC
R5 6-13, 68161 Mannheim, Germany / Drift 10, 3512 BS Utrecht, The Netherlands
{kamocki | witt}@ids-mannheim.de

Abstract

Privacy by Design (also referred to as Data Protection by Design) is an approach in which solutions and mechanisms addressing privacy and data protection are embedded through the entire project lifecycle, from the early design stage, rather than just added as an additional layer to the final product. Formulated in the 1990 by the Privacy Commissionner of Ontario, the principle of Privacy by Design has been discussed by institutions and policymakers on both sides of the Atlantic, and mentioned already in the 1995 EU Data Protection Directive (95/46/EC). More recently, Privacy by Design was introduced as one of the requirements of the General Data Protection Regulation (GDPR), obliging data controllers to define and adopt, already at the conception phase, appropriate measures and safeguards to implement data protection principles and protect the rights of the data subject. Failing to meet this obligation may result in a hefty fine, as it was the case in the Uniontrad decision by the French Data Protection Authority (CNIL). The ambition of the proposed paper is to analyse the practical meaning of Privacy by Design in the context of Language Resources, and propose measures and safeguards that can be implemented by the community to ensure respect of this principle.

Keywords: GDPR, personal data, Privacy by Design

1. Introduction. Presentation of the concept of Privacy by Design

The principle of Privacy by Design has drawn considerable attention of the public since the adoption of the EU's General Data Protection Regulation (GDPR) in 2016, and especially since its entry into force in 2018. One should not forget, however, that, unlike the GDPR, this principle has in fact been around for decades, and that it originated outside of Europe.

Although some sources would trace the origins of the concept back to the 1970s, the paternity of privacy by design is commonly attributed to Ann Cavoukian, the Privacy Commissionner of Ontario, Canada, co-author of the 1995 international report on Privacy Enhancing Technologies (PET). Cavoukian famously argued "that the future of privacy cannot be assured solely by compliance with regulatory frameworks; rather, privacy assurance must ideally become an organization's default mode of operation" (Cavoukian, 2009). In this approach, to use a common metaphor, privacy protection should be 'baked in' a technology or a product, rather than just sprinkled over it.

More recently, in 2009, Cavoukian listed what she called 7 Foundational Principles of Privacy by Design:

- Proactive not reactive, Preventative, not Remedial
- Privacy as the default
- Privacy Embedded into Design
- Full functionality - Positive Sum not Zero Sum
- End-to-end security - Lifecyle Protection
- Visibility and Transparency
- Respect for User Privacy

By the time these rules were formulated, the imperative of embedding privacy into design was already sanctioned in EU legislation, albeit rather timidly: Recital 46 of the 1995 Data Protection Directive (95/46/EC) stated that 'the protection of (...) personal data requires that appropriate technical and organizational measures be taken, both at the time of the design of the processing system and at the time of the processing itself'. In the past decade, however, the concept gained much more traction on both sides of the Atlantic.

In 2010, Privacy by Design was mentioned in a major EU policy document, Digital Agenda for Europe (COM(2010)245), as essential for practical enforcement of the right of privacy and to the protection of personal data in the EU. The Commission defined Privacy by Design as an approach in which "privacy and data protection are embedded throughout the entire life cycle of technologies, from the early design stage to their deployment, use and ultimate disposal". This definition was also mentioned in a 2010 Commission communication entitled "A comprehensive approach on personal data protection in the European Union" (COM(2010) 609).

Privacy by Design was also discussed by US policymakers. In 2012, the Federal Trade Commission adopted its Recommendations for Businesses and Policymakers concerning Protecting Consumer Privacy in an Era of Rapid Change (FTC, 2012). The document listed Privacy by Design (defined as an approach in which "companies should promote consumer privacy throughout their organizations and at every stage of the development of their products and services") as one of the key recommendations.

Arguably the heyday of Privacy by Design is yet to come. As mentioned above, Privacy by Design is now an important part of the EU's General Data Protection Regulation (adopted in 2016 and entered into force on 25 May 2018). Article 25(1) of the Regulation provides that "Taking 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, the controller[1] shall, both at the time of the deter-

[1] Article 4(7) of the GDPR defines the controller as 'natural or legal person, public authority, agency or other body which, alone or jointly with others, determines

mination of the means for processing and at the time of the processing itself, implement appropriate technical and organisational measures, such as pseudonymisation, which are designed to implement data-protection principles, such as data minimisation, in an effective manner and to integrate the necessary safeguards into the processing in order to meet the requirements of this Regulation and protect the rights of data subjects".

Despite the criticism for its tautological wording that Privacy by Design received from some representatives of the academic community (Gürses et al., 2011), its disrespect is now -- at least in the EU -- a potential ground for a hefty fine. Quite recently, the French Data Protection Authority fined a small translation company (Uniontrad) 20 000 EUR for failing to observe Privacy by Design[2].

Aware of the rather theoretical character of this principle, data protection authorities have started adopting guidelines concerning Privacy by Design, thereby making it more concrete and tangible. Such guidelines were adopted by the European Data Protection Supervisor (EDPS, 2018), the Spanish Data Protection Authority (AEPD, 2019) and very recently also by the European Data Protection Board (EDPB, 2019). The ambition of this paper is to briefly present the conclusions of these reports and discuss their relevance for the Language Resources (LR) community.

2. The Importance of Privacy by Design for the LR community and its role in public tenders

Just like in any data-intensive discipline, the LR community is all about data. It is a well-known fact that LRs often contain information relating to identified or identifiable persons (i.e. personal data[3]), such as names, dates of birth, information on personal background, education, preferences, professional life, or in some cases recordings featuring the person. The simplistic approach according to which any data found on the Internet is public, and not personal, and therefore can be freely reused for LRs is thankfully fading away, but some misconceptions about the role of data protection in compiling LRs are still present to this day[4]. Some members of the community are still ready to argue e.g. that data related to the professional sphere of a person's activity should not be regarded as 'personal data', or that 'the purpose is not to spy on people, but to build LR, therefore GDPR does not apply'. Many still seem to believe that any data protection issues are simply handled by an attempt on anonymisation just before the LR is made publicly available, if ever.

Such practice is in stark contrast with the fact that many LR projects are financed by the public sector, or even by the European Commission itself. In this context, it should be kept in mind that Recital 78 of the General Data Protection Regulation clearly states that Privacy by Design should be taken into consideration in the context of public tenders. The implementation of this principle should therefore be seen not as another burden, but as an opportunity to gain competitive advantage and increase funding opportunities.

With this in mind, the following section will explore the practical meaning of Privacy by Design.

3. Implementing Privacy by Design in Practice

Privacy by Design has sometimes been overlooked due to its unclear scope and meaning. It is therefore of paramount importance to understand what is really expected from data controllers under this principle, especially in light of recent guidelines published by various data protection bodies.

In essence, Article 25(1) of the GDPR (quoted above) obliges data controllers to (1) implement appropriate technical and organisational measures to implement the data protection principles (2) integrate the necessary safeguards into the processing in order to meet the requirements of the GDPR and protect the rights of data subjects.

3.1 Technical and organisational measures

According the the EDPB, 'measures' should be understood as 'any method or means' employed during the processing. These measures can be technical (e.g. pseudonymisation) or organisational (e.g. training sessions for personnel who participates in personal data processing operations).

These measures should be adopted to implement the data protection principles set forth in the GDPR. It is therefore crucial to understand what these principles are.

According to the list in Article 5 of the GDPR, the following principles apply to data processing:

- lawfulness (processing should always have an appropriate legal basis, such as the data subject's consent or legitimate interest of the controller);
- fairness;
- transparency (information regarding e.g. the purposes of processing, the identity of the controller and data rendition periods should be made available to the data subjects);

the purposes and means of the processing of personal data'.

[2] CNIL, Délibération de la formation restreinte n° SAN-2019-006 du 13 juin 2019 prononçant une sanction à l'encontre de la société UNIONTRAD COMPANY.

[3] Article 4(1) of the GDPR defines personal data as 'any information relating to an identified or identifiable natural person ('data subject')'.

[4] For an example of this attitude, see a LREC Helpdesk ticket : http://helpdesk.lr-coordination.eu/view/343/.

- purpose limitation (data should be processed for a specific purpose and not further processed for purposes incompatible with the original purpose);
- data minimisation (data should be “adequate, relevant and limited to what is necessary in relation to the purposes” for which they are processed);
- accuracy (data should be accurate, and where necessary kept up to date; inaccurate data should be erased or rectified);
- storage limitation (data can only be kept for no longer than necessary to achieve the purposes of processing; the retention period, or at least the criteria used to determine it, should be defined and communicated to the data subject);
- integrity and confidentiality (data should be processed in a manner that ensures appropriate technical and organisational security);
- accountability (the controller should be able to demonstrate compliance with the GDPR).

Due to the limited space in this article, it is impossible to discuss each and every one of these principles (for a short overview, see e.g. Kamocki et al., 2018a) and appropriate measures relating to them in great detail. Instead, we would like to focus on two principles that seem to be the most problematic from the point of view of LRs, namely data minimisation and storage limitation. It should be noted that while the storage limitation principle may be derogated from when data are processed exclusively for research purposes (if appropriate safeguards such as pseudonymisation are implemented, cf. Art. 89 of the GDPR), the data minimisation principle goes with no such exceptions. For a more detailed discussion on Privacy by Design in the field of Machine Translation, see Kamocki, Stauch, 2020.

It is no time and place to discuss the justifiability of data protection principles (which, we agree, may seem questionable from the point of view of data-intensive science and technology); regardless of whether we find them ‘good’ and ‘reasonable’ or not, GDPR principles are now binding on all those who process personal data, and their disrespect may result in a hefty fine of up to 20 000 000 EUR, together with other adverse consequences related to bad publicity and loss of trust. It is therefore timely and useful to discuss how these principles should be put into action.

3.1.1 Data minimisation

Regarding data minimisation, measures related to this principle should focus on assessing necessity of the collected data and avoiding collection of unnecessary data. For example, if data are collected *via* web crawling, as it is often the case in LR projects, the danger of processing (at least collecting and storing) unnecessary (i.e. excessive) personal data is substantial. In order to mitigate this risk, the crawler could be trained to avoid personal data altogether, or to automatically anonymise or pseudonymise the data instantly upon collection. The former can be achieved by avoiding certain categories of websites likely to contain personal data (such as social media, personal blogs etc.) or by avoiding to scrape named entities or sequences of data likely to contain postal or e-mail addresses, or phone numbers. The latter can be aimed at by training the crawler to automatically anonymise or pseudonymise the data upon its collection. Randomisation, a technique in which the data are shuffled within the same category (adresses with adresses, names with names etc.) may be a good approach in many cases, as it allows to (at least partially) preserve the linguistic structure of the data. In speech LRs, specific privacy-protecting techniques applicable to speaker characterisation and speech characterisation, such as those described by Nautsch et al. (2019), can be implemented.

3.1.2 Storage limitation

Regarding the principle of storage limitation, it can be implemented via measures such as adopting and following a clear policy listing the criteria used to determine the retention periods, or automating the processes of anonymisation, archiving or deletion. It may be useful, for example, to design a tool that would periodically check if the collected personal data in a LR are still available at their original source, which would also allow to determine the accuracy of the data. Data that were deleted or rectified in the original source should be considered for deletion from the LR. Even simple and low-cost methods, such as setting alerts related to data retention periods, may be effective in implementing the storage limitation principle. In any case, it is crucial to remember that GDPR in principle prohibits infinite storage of personal data, unless a specific legal provision (e.g. related to public archives) applies. When the data are processed for research purposes with appropriate safeguards (such as pseudonymisation), some leniency (storage for longer than necessary) is also allowed (Art. 5.1 (e) of the GDPR).

3.2 Safeguards

Apart from organisational and technical measures, controllers should also adopt safeguards to ensure respect of the abovementioned principles, and to protect the rights of data subjects. For the most part, the rights of data subjects are listed in Articles 12 to 22 of the GDPR, and include:

- information;
- withdrawal of consent (where processing is based on consent);
- access;
- rectification;
- erasure (‘right to be forgotten’);
- data portability;
- right to object;
- freedom from automated decision-making, including profiling.

These rights can be safeguarded by adopting specific policies and procedures concerning the ways in which information is provided to data subjects, and how their requests regarding the exercise of their rights (e.g. access and rectification) are handled. Such procedures should define contact points for data subjects to exercise their rights, and how requests and complaints are further processed within the organisation. The interest of having such policies and procedures resides mostly in clear definition of responsibilities. It also theoretically allows to avoid problems related to personal rotations within the team. The policies and procedures should also define specific deadlines for taking required actions.

Appropriate safeguards should also be implemented in order to meet other requirements of the GDPR. This seem to apply especially to such obligations as security (Article 32) and notification and/or handling data breaches (Articles 33 and 34), which can be safeguarded by adopting appropriate Data Breach Policies. Just like the policies and procedures discussed above and concerning the exercise of data subject rights, a Data Breach Policy should determine who should be notified in case of an incident which may potentially constitute a data breach (e.g. loss of a USB stick), who will define actions aimed at containing the breach, and how it will be determined whether the breach should be notified to the competent data protection authority and/or communicated to data subjects.

3.3 Effectiveness of measures and safeguards

In order to meet the Privacy by Design requirement, measures and safeguards should be adopted already in the conception phase (the time of determination of purposes and means), and taken into account throughout the whole processing stage (up until erasure or anonymisation of the data). Such measures, however, need to be effective; according to the EDPB, controllers should not simply adopt generic measures and safeguards, but instead justify the choice of each measure and document its actual effect for the particular processing. In other words, a simple measure of shuffling named entities and dates may be efficient for some LRs, but quite useless (from the privacy standpoint) for others. Before a measure is implemented, its efficiency should be assessed (in a documented manner). Efficiency of technical measures can often be assessed qualitatively, whereas efficiency of organisational measures (policies) can be assessed in practice, or by conducting drills.

3.4 Elements to be taken into account

According to the GDPR, in choosing appropriate measures and safeguards, the following elements should be taken into account by the controller:

- state of the art (which means that as technology evolves, the choice of measures and safeguards should be re-evaluated);
- cost of implementation (EDPB clearly states that on the one hand spending more does not necessarily lead to better results, while on the other hand incapacity to bear implementation costs is no excuse for lack of compliance with the GDPR)
- nature, scope, context and purpose of the processing; and
- risks of varying likelihood and severity for rights and freedoms of natural persons posed by the processing.

In some LR projects, the two last factors can be used to argue that -- due to the nature of the processing and low risks for individuals -- data protection can be sufficiently guaranteed with relatively basic measures; however, total absence of measures and safeguards can in no case be justified.

4. Role of certification mechanisms

Article 25(3) of the GDPR clearly states that an approved certification mechanism may be used as an element to demonstrate compliance with Privacy by Design. Therefore, obtaining a certificate, such as the European Data Protection Seal, can make it easier for controllers to demonstrate compliance with Privacy by Design (among other principles of the GDPR). Although not expressly provided for in the GDPR, in our opinion the same can possibly be said about adherence to an approved Code of Conduct (such as the Code of Conduct currently under discussion within the CLARIN community (Kamocki et al., 2018b)), especially if the Code of Conduct expressly adresses the question of appropriate technical and organisational measures and safeguards. This is yet another reason to consolidate the LR community around the idea of a GDPR Code of Conduct.

5. Bibliographical References

Agencia Espanola Proteccion Datos (AEPD) (2019). A Guide to Privacy by Design.

Cavoukian, A. (2009), Privacy by Design. The 7 Foundational Principles. Available ar: https://www.ipc.on.ca/wp-content/uploads/Resources/7foundationalprinciples.pdf.

European Data Protection Board (EDPB) (2019). Guidelines 4/2019 on Article 25 Data Protection by Design and by Default.

European Data Protection Supervisor (EDPS) (2018). Opinion 5/2018. Preliminary Opinion on privacy by design.

Gurses, S., Troncoso, C. and Diaz, C. (2011). Engineering Privqcy by Design. Computers, Privacy & Data Protection (CPDP).

Information and Privacy Commissioner/Ontario, and Registratiekamer (1995). Privacy-Enhancing Technologies. Volume 1.

Kamocki, P., Ketzan, E. and Wildgans J. (2018a). Language Resources and Research under the General Data Protection Regulation, CLARIN White Paper Series.

Kamocki, P., Ketzan, E., Wildgans, J. and Witt, A. (2018b) Toward a CLARIN Data Protection Code of Conduct [in:] Skadinq, I., Eskevich, M. (Eds.) : CLARIN Annual Conference 2018, Proceedings. 8-10 October 2018, Pisa, Italy.

Kamocki, P., Stauch, M. (2020), "Cover this data that I cannot see". Privacy by Design in Machine Translation, [in:] Prosiel, J. (Ed.), Maschinelle Übersetzung für Übersetzungsprofis, BDÜ.

Nautsch , A.et al., (2019). Preserving privacy in speaker and speech characterisation, Computer Speech & Language 58 (2019) 441-480.

Strötgen, J. and Gertz, M. (2012). Temporal tagging on different domains: Challenges, strategies, and gold standards. In Nicoletta Calzolari (Conference Chair), et al., editors, Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), pages 3746–3753, Istanbul, Turkey, may. European Language Resource Association (ELRA).