Data Management Plan (DMP) for Language Data under the New General Data Protection Regulation (GDPR)

Pawel Kamocki, Valérie Mapelli, Khalid Choukri
ELDA
9 rue des Cordelières, 75013 Paris
{kamocki | mapelli | choukri} @elda.org

Abstract
The EU's General Data Protection Regulation (GDPR) of 27 April 2016 will apply from 25 May 2018. It will reinforce certain principles related to the processing of personal data, which will also affect many projects in the field of Natural Language Processing. Perhaps most importantly, the GDPR will introduce the principle of accountability, according to which the data processor shall be able to demonstrate compliance with the new rules, and that he applies 'privacy by design and by default'. In our opinion, a well-drafted Data Management Plan (DMP) is of key importance for GDPR compliance; indeed, the trend towards the adoption of a DMP, particularly in EU-funded research projects, has been more vivid since 2017, after the extension of the Horizon 2020 Open Data Pilot. Since 2015, ELRA also proposes its own template for the Data Management Plan, which is being updated to take the new law into account. In this paper, we present the new legal framework introduced by the GDPR and propose how the new rules can be integrated in the DMP in order to increase transparency of processing, facilitate demonstration of GDPR compliance and spread good practices within the community.

Keywords: data management plan, data protection, anonymisation, personal data

1. Introduction

In the years 2017-2018, the language resources community, especially in the European Union, has been confronted with certain important changes. First of all, the Regulation 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 (General Data Protection Regulation, hereinafter: GDPR) will apply from 25 May 2018. This new legislative act will repeal the Directive 95/46/EC and unify data protection laws across the whole European Union. Unlike a directive, a regulation applies directly in all the Member States, it is therefore essential for researchers and companies established in the EU[1] and processing language data to comply with the GDPR.

Secondly, the trend towards the adoption of Data Management Plans has recently been more visible. This is partly due to the extension of the Open Research Data Pilot to all the thematic areas of the Horizon 2020 Programme (hereinafter: H2020). Indeed, from 2017 on, all the research data in the H2020-funded projects is open by default (with opt-outs still possible); these projects are therefore obliged to adopt a Data Management Plan (DMP) addressing the issues of Findability, Accessibility, Interoperability and Reusability (FAIR) of the data (see art. 29.3 of the H2020 Model Grant Agreement). A template for a DMP is available on the European Commission's web portal[2]. Of course, the importance of a DMP is not limited to H2020-funded projects — in fact, many other projects and institutions have adopted DMPs long before the extension of the Open Research Data Pilot. For example in the US, the National Science Foundation (NSF) requires all grant proposals to include a DMP 'of no more than two pages'.

Since 2015, ELRA has proposed a model for a DMP concerning specifically language resources (Choukri et al., 2016).

The purpose of the proposed paper is to draw the attention of the community on how certain issues related to the processing of personal data should be integrated in a DMP to ensure compliance with the GDPR. It is therefore useful to examine the general framework of the GDPR and then to see how the DMPs should be modified in order to adapt to the new law.

2. General Data Protection Regulation – a summary of the new framework

Most of the changes introduced by the GDPR are of evolutionary rather than revolutionary nature; the light has been shed on some grey areas, but the main principles remain largely the same (2.1). The most important addition from the point of view of the language community is probably the reinforcement of the obligations related to data processing (2.2).

2.1. General framework of the GDPR – what's old, what's new?

The notions of personal data and processing remain unchanged for the most part. Personal data are defined, just like in the Directive 95/46/EC, as *"any information relating to an identified or identifiable natural person"* (art. 4 no. 1 of the GDPR). According to the opinion of the Article 29 Data Protection Working Party (hereinafter: WP29)[3], the notion covers not only 'objective' information (i.e. facts), but also 'subjective' information (opinions and assessments). The information 'relates to a person' not only if it is 'about' a person (the 'content' element), but also if it is used to evaluate or influence the status or behaviour of the person (the 'purpose' element), or if it has an impact on

[1] …and not only; in fact, the GDPR applies also to entities established outside of the EU if they offer goods and services to physical persons in the Union, or if they monitor the behavior of physical persons on the Union's territory (art. 3 of the GDPR).

[2] http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf

[3] WP29 is a working group made up of a representative of the data protection authority of each EU Member State, the European Data Protection Supervisor and the European Commission, created by art. 29 of the Directive 95/46/EC. Under the GDPR, it will be replaced by the European Data Protection Board.

the person's interests or rights (the 'result' element) (WP29, 2007). For the definition of personal data to be met, the person that the information relates to may be identified (i.e. singled out of a group), but also identifiable (i.e. possible to be singled out) directly (e.g. by a name or by an identification number) or indirectly (e.g. by a unique combination of various factors, such as sex, date of birth and postal code) (WP29, 2007). The concept of personal data is therefore extremely broad; it refers not only, as some may believe, to data containing named entities. In particular, all the unaltered video and voice recordings involving physical persons shall be regarded as personal data, as they often allow to identify the speaker. Indeed, recital 26 of the GDPR specifies that in order to determine whether a natural person is identifiable *"account should be taken of all the means reasonably likely to be used".* The recital continues: *"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".* It is therefore expressly stated that identifiability may change over time – data that do not allow to identify the data subject today, may identify him tomorrow. Consequently, it is necessary to periodically review the results of anonymization processes. On the other hand, data that do not relate to any natural person or that do not allow to identify the natural person that they relate to (because they have been successfully anonymised) shall be regarded as anonymous data and can be freely processed.

The notion of processing is also very broad; it is defined as *"any operation or set of operations which is performed on personal data (...) whether or not by automated means"* (art. 4 no. 2 of the GDPR). In particular, this includes collection, storage, consultation, adaptation, but also erasure. Personal data have to be processed in a way that is lawful (see below), fair and transparent (art. 5.1(a) of the GDPR). The data can only be collected for specified, explicit and legitimate purposes, and not further processed in a manner incompatible with these purposes (purpose limitation — art. 5.1(b) of the GDPR). The data that are being processed have to be limited to what is necessary in relation to the purposes of processing (data minimization – art. 5.1(c) of the GDPR) and only kept for as long as necessary (storage limitation — art. 5.1, (e) of the GDPR). They should be accurate and, when necessary, kept up to date (art. 5.1(d) of the GDPR). Moreover, they should be processed in a manner that ensures appropriate security, *"including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures"* (art. 5.1(f) of the GDPR).

The processing is lawful if one of the conditions set forth in art. 6 of the GDPR is met. Chiefly, it is the case when the data subject (i.e. the person that the data refer to) has given his consent to the processing (art. 6.1(a) of the GDPR). The GDPR contains some important precisions when it comes to the conditions for valid consent (see esp. art. 4 no. 11 of the GDPR): most importantly, it should be freely given, specific (i.e. limited to a specific purpose of processing), informed (i.e. prior to giving his consent, the data subject should be informed at least about the identity of the data controller and the purposes of the processing – recital 42 of the GDPR) and unambiguous. It can be withdrawn by the data subject at any moment (art. 7.3 of the GDPR). On the other hand, consent does not necessarily have to be given in a written document – it can also be an oral statement or any other unambiguous indication of the data subject's agreement to the processing[4].

Apart from consent, other legal grounds for processing are also possible; from the perspective of the language resources community, the most important of these alternative grounds is the pursuit of legitimate interests of the controller (arguably, research can be such a legitimate interest). Processing can indeed be based on this ground, unless the legitimate interests of the controller are overridden by the interests or fundamental rights and freedoms of the data subject (art. 6.1(f) of the GDPR). In assessing this, account should be taken of the reasonable expectations of the data subjects, as well as the applied safeguards (WP29, 2014).

This brings us to the last point of this paragraph: one of the new and interesting additions in the GDPR which is the introduction of the notion of pseudonymisation, absent from the Directive 95/46/EC. Pseudonymisation is defined as *"the 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"* (art. 4 no. 5 of the GDPR). Pseudonymisation is specifically not equivalent to anonymization (pseudonymised data are still to be regarded as personal data – recital 26 of the GDPR), but it is a safeguard measure that can be taken into account in assessing the risks of processing for the data subject.

2.2. Reinforced obligations of the data controller

The data controller is an entity that determines the purposes and means of processing of personal data. The GDPR considerably reinforces the obligations of the data controller.

Most importantly, the GDPR introduces the principle of accountability, according to which the controller shall be responsible for compliance of the processing with the principles of the GDPR, and able to demonstrate this compliance (art. 5.2 of the GDPR). In particular, whenever processing is based on the data subject's consent, the controller should be able to demonstrate this consent (art. 7.1 of the GDPR). It is also his responsibility to implement appropriate technical and organisational measures and, when necessary, policies in order to ensure compliance with the GDPR (art. 24 of the GDPR).

Moreover, such technical and organizational measures should be implemented *"both at the time of the determination of the means for processing and at the time of the processing itself".* This is particularly important to ensure that the principle of data minimization (see above) is respected. Therefore, the data controller is obliged to implement

[4] For more information about consent, see: WP29, 2017.

"data protection by design and by default" (art. 25 of the GDPR).

The controller is also (with some exceptions) obliged to maintain a written record of his processing activities (see art. 30 of the GDPR for details). In addition to that, when processing presents a high risk to rights and freedoms of natural persons, the controller is obliged to carry out, prior to the processing, an assessment of the impact of the envisaged processing (art. 35 of the GDPR) (WP29, 2017a).

In order to comply with the transparency principle, the controller should also adopt appropriate measures to provide the data subject with information listed in articles 13 and 14 of the GDPR, and in particular the identity of the controller, the purpose of the processing, the recipients of envisaged data transfers, the period for which the data will be stored and the rights of the data subject with regards to his personal data (see art. 15 through 21 of the GDPR).

3. Data Management Plan in H2020

The H2020 DMP Template consists of six sections: (1) Data Summary, (2) FAIR (Findable, Accessible, Interoperable and Re-Usable) Data, (3) Allocation of Resources, (4) Data Security, (5) Ethical aspects, (6) Other.

Since under the GDPR data processors have to implement the *"privacy by design and by default"* approach, and shall be able to demonstrate their compliance with this and other GDPR rules according to the accountability principle (see above), in the projects involving processing of personal data it is useful to integrate certain principles of the GDPR in the DMP, even using the exact wording of the GDPR. While this is not by itself sufficient to prove actual compliance with the GDPR, it would undoubtedly help demonstrate the processor's proactive attitude and spread good practices within the community. A well-drafted (in plain, understandable language) DMP should also be made available to data subjects in order to comply with the transparency requirement (see above). Moreover, it is an excellent basis for the record of personal data processing activities which processors of such data are obliged to keep.

We will now see how different GDPR rules can be incorporated into various sections of the DMP.

3.1. Data Summary

The Data Summary is an introductory section which, among others, states the purpose of data collection. In our opinion, it is useful not only to clearly state the purpose of processing in that section (according to the principle of purpose limitation of art. 5.1(b) of the GDPR), but also to mention the principles of data minimisation (art. 5.1(c) of the GDPR) and data accuracy (art. 5.1(d) of the GDPR), as well as on which legal ground personal data are being processed (i.e. consent or some alternative ground).

3.2. Data Accessibility

Section 2.2 of the H2020 DMP Template addresses data accessibility. It is this section that is supposed to explain how and to whom the data is going to be made available. In the terminology of the GDPR this corresponds to the *"envisaged transfers"* of data, of which the data processor have to inform the data subject. Moreover, the data subject has the right to access the data (according to art. 15 of the GDPR), which should also, in our opinion, be stated in this section of the DMP.

3.3. Data Re-Usability

Section 2.4 of the H2020 DMP Template concerns the reusability of the data. While this section normally deals with IPR licensing questions, in our view it is important to mention three aspects related to data protection (and which also have to do with data security): anonymisation, storage limitation (art. 5.1(e) of the GDPR) and purpose limitation (art. 5.1(b) of the GDPR). Indeed, it is useful to state in this section that personal data will not be re-used for purposes incompatible with the purpose for which they were initially collected (as stated in the Data Summary section) and stored for longer than necessary to achieve this purpose. As soon as the purpose can be achieved without processing personal data, the data shall be anonymised.

3.4. Allocation of Resources

Section 3 of the H2020 DMP Template addresses *"Allocation of Resources"*. While the primary concern of this section is the estimation and coverage of financial costs of processing, it should also *"clearly identify responsibilities for data management in [the] project"* (H2020, 2016). Therefore, this is where the data controller should be clearly identified (the entity that defines the means and purposes of data processing), and data processors (entities processing the data on behalf of the controller) should be named. It is also useful to restate some of the rules regarding the relation between the data controller and data processors (art. 28 of the GDPR), such as the one according to which the processor *"processes personal data only on documented instructions from the controller"* and *"shall not engage another processor without prior specific or general written authorisation of the controller"*.

3.5. Data Security

Section 4 of the H2020 DMP Template is fundamental from the point of view of GDPR-compliance. Indeed, the principle of data integrity and confidentiality (art. 5.1(f) of the GDPR) requires that data shall be processed *"in a manner that ensures appropriate security (...), including protection against unauthorised or unlawful processing and against accidental loss, destruction or damage, using appropriate technical or organisational measures"*. Apart from specific security measures adopted in the project (such as pseudonymisation — see art. 32 of the GDPR), this section should also expressly state that the data are processed according to this general principle.

3.6. Ethical Issues

Section 5 of the H2020 DMP Template is the appropriate place to mention the results of a data protection impact assessment (art. 35 of the GDPR) if such an assessment has been carried out. Moreover, it is useful to mention in this section that such an assessment will be carried out whenever required by law. In our opinion, this section is also a good place to mention that any security breaches will be promptly reported to the national supervisory authority (as required by art. 33 of the GDPR) and communicated to the data subjects (as per art. 34 of the GDPR).

4. The role of ELRA/ELDA

In order to help the language resources community adapt to the new rules, ELRA will review its DMP. While distributing language resources, ELDA will also act as an intermediary between data subjects and the data controller. Moreover, it also proposes a legal helpdesk[5], freely available to those who want to learn more about the new Regulation.

4.1. ELRA Data Management Plan

Since 2015, ELRA proposes its own DMP template, designed specifically with language resources in mind (Choukri et al., 2016). Recently this DMP has been reviewed and adapted to the GDPR. The "Project Description" section now includes a statement on whether personal data are processed within the project. If the answer is in the positive, other elements of the DMP are modified accordingly. In particular:

○ the "Project Description" section:

- clearly and unambiguously identifies the data controller (or joint controllers);
- contains a statement that if a data processor is involved, he shall *« processes personal data only on documented instructions from the controller »* and *« shall not engage another processor without prior specific or general written authorisation of the controller »*;
- states the purpose for which personal data are processed and restates the principle of purpose limitation (art. 5.1 (b) of the GDPR);

○ the "Data Acquisition" section:

- restates the principles of data minimisation (art. 5.1 (c) of the GDPR) and data accuracy (art. 5.1 (d) of the GDPR);
- clearly identifies the legal ground for processing of personal data (consent or an alternative ground) and restates the principle of lawfulness (art. 6 of the GDPR);

○ the "Legal Issues and Ethics" section:

- restates the obligation to anonymise personal data as soon as possible in relation to the purposes of processing;
- restates the obligation to carry out a Data Protection Impact Assessment and presents the results of the DPIA if it was carried out;

○ the "Sustainability" section:

- reminds that according to the principle of data accuracy (art. 5.1 (personal data have to be kept up to date;
- reminds that the results of anonymization shall be periodically reviewed to take into account technological progress in identification techniques;

○ the "Data Storage Section" restates the principle of data integrity and confidentiality (art. 5.1 (f) of the GDPR), and the principle of storage limitation (art. 5.1 (e) of the GDPR);

○ the "Data Access and Sharing" section:

- lists envisaged data transfers;
- reminds of the data subject's right of access (art. 15 of the GDPR).

These changes are intended to educate the community about the current legal framework and to spread good practices. The new ELRA DMP will also – if applied and followed by the users – facilitate the demonstration of compliance with the GDPR, required under the principle of accountability (art. 5.2 of the GDPR, see above).

4.2. ELDA as an intermediary

While distributing language resources, ELDA will also play the role of an intermediary between the data subject and the data controller. This role will consist primarily of providing the data subjects with information required by the GDPR (art. 13 and 14 of the GDPR), according to the transparency principle. ELDA will also guarantee the data subject's right to withdraw his consent to further processing of his personal data (as well as the right to erasure, restriction and objection) by enabling him to communicate his wishes directly to ELDA which will then transmit it to the controller.

4.3. ELDA's Legal Helpdesk

ELDA's Legal Helpdesk will also provide, free of charge, any member of the community with information regarding the GDPR and recommendations on how to comply with the new rules. Moreover, ELDA can also offer assistance (including on-site) regarding data protection and data management.

5. Closing remark — the role of the Codes of Conduct

In the previous sections, we have established that a well-drafted DMP is essential for demonstrating compliance with the GDPR, and in particular the principles of transparency and accountability. However, it is not the only tool to demonstrate compliance with the GDPR. Another such tool would be a community-wide Code of Conduct regarding processing of personal data for Natural Language Processing purposes. Indeed, art. 40 of the GDPR encourages the adoption of Codes of Conduct by *"associations and other bodies representing categories of controllers or processors"*. Such codes have to be approved, registered and published by the national supervisory authority; it may be even granted universal validity by the European Commission (and "complement" the GDPR). The GDPR also specifies the process of monitoring compliance with such codes of conduct (art. 41), and even certification processes (art.

[5] http://www.elra.info/en/services-around-lrs/legal-support-helpdesk/

42). The adoption of such a code of conduct would require a substantial, internationally coordinated effort on behalf of the language resources community, but the reward (simplified demonstration of compliance and lower transaction costs) may be well worth it.

If such an international Code of Conduct is adopted, or if national codes are adopted at the level of some Member States, the adherence to such a Code should also be mentioned in the DMP.

6. Bibliographical References

Choukri, K.; Mapelli, V.; Mazo, H.; Popescu V. (2016). ELRA Activities and Services. In: Proceedings of LREC 2016, pp. 463-468.

H2020 Programme (2016). Guidelines on FAIR Data Management in Horizon 2020, version 3.0, 26 July 2016.

H2020 Programme (2017). Annotated Model Grant Agreement, Version 4.0.1, 20 June 2017.

Regulation 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.

WP29 (Article 29 Data Protection Working Party) (2007). Opinion 4/2007 on the concept of personal data.

WP29 (Article 29 Data Protection Working Party) (2014). Opinion 06/2014 on the notion of legitimate interests of the data controller under Article 7 of Directive 95/46/EC.

WP29 (Article 29 Data Protection Working Party) (2017). Guidelines on Consent under Regulation 2016/679.

WP29 (Article 29 Data Protection Working Party) (2017a). 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.